

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

Aashish Dhawan*

University of Florida

aashish.dhawan@ufl.edu

Christopher Driggers-Ellis*

University of Florida

driggersellis.cw@ufl.edu

Dzmitry Kasinets*

University of Florida

dkasinet@ufl.edu

Christan Grant

University of Florida

christan@ufl.edu

Daisy Zhe Wang

University of Florida

daisyw@cise.ufl.edu

Abstract

This paper describes the University of Florida gators submission to the WMT26 Low-Resource Indic Language Translation shared task. We adapt the retrieval-augmented many-shot translation pipeline from our AmericasNLP 2026 system (Dhawan et al., 2026b; Bui et al., 2026) to translate between English and eleven North-Eastern Indian languages in both directions. At inference time, BM25 (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) translates the input conditioned on these examples. No model fine-tuning is involved. Training banks combine official WMT26 data with publicly available corpora such as Samanantar (Ramesh et al., 2022) and prior WMT shared task releases. 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. Find our code on [Github](#).

1 Introduction

The WMT26 Low-Resource Indic Language Translation shared task (Pakray et al., 2026) covers eleven North-Eastern Indian languages that span four language families: Assamese (Indo-Aryan), Khasi (Austroasiatic), eight Tibeto-Burman languages (Mizo, Manipuri in both Bengali and Meitei Mayek scripts, Bodo, Kokborok, Karbi, Nyishi, and Tagin), and Nagamese, an English-based creole. Several of these — particularly Nyishi (Kakum et al., 2023), Karbi, and Tagin — have almost no prior MT work.

The system we describe here emerges from two earlier projects. In our LoResMT 2026 pa-

per (Dhawan et al., 2026a), we found that augmenting mBART with NLLB-200-generated synthetic data improved ChrF++ for low-resource Indigenous American languages, but the gains were modest and domain-sensitive. We then moved to a retrieval-augmented approach for the AmericasNLP 2026 captioning task (Dhawan et al., 2026b; Bui et al., 2026): instead of fine-tuning, we used BM25 to pull relevant parallel examples into a long Gemini 2.5 Flash (Comanici et al., 2025) prompt and let the model translate in context. That system won the overall shared task competition across five Indigenous American languages, and the gap over fine-tuned mBART was large enough that we wanted to test whether the approach would transfer to a different language family and task domain.

For WMT26, the adaptation is straightforward. The AmericasNLP pipeline had two stages — VLM image captioning in Spanish, then retrieval-augmented Spanish-to-LoRes translation. Since WMT input is already English text, we drop the vision stage and feed source sentences straight into the BM25+Gemini pipeline. The rest carries over: per-language system prompts, the r, d hyperparameter sweep, and the same scoring infrastructure.

2 Background

The WMT Low-Resource Indic shared task has run annually since 2023 (Pal et al., 2023), growing from four language pairs to eleven in 2026 (Pakray et al., 2026). The task splits languages into two categories: Category 1 (moderate data) includes Assamese, Mizo, Khasi, Manipuri in Bengali script, Manipuri in Meitei Mayek, and Nyishi; Category 2 (very limited data) includes Bodo, Kokborok, Karbi, Nagamese, and Tagin. The Indic shared task organizers produce no official ranking

**Equal contribution.

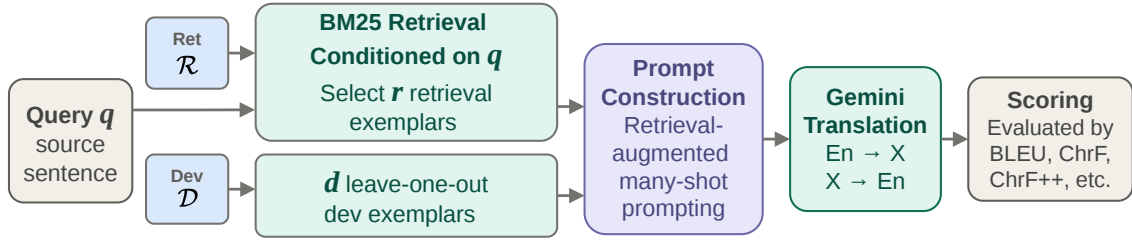


Figure 1: Overview of the **gators** pipeline. The query q is a sentence in the source language. We condition retrieval from the set of exemplars $\mathcal{R}$ on q and pull d exemplars from $\mathcal{D}$. The r and d exemplars are placed with q into the Prompt Construction pipeline for translation by Gemini Flash 2.5 into English (En) or Indic language X.

of submissions, nor is a single winner declared, but they do report numerous metrics including BLEU (Papineni et al., 2002), ChrF++ (Popović, 2017), COMET (Rei et al., 2020), and others.

Previous editions (Pakray et al., 2024, 2025) have been dominated by fine-tuned multilingual models — IndicTrans2 (Gala et al., 2023), mBART (Liu et al., 2020), and NLLB-200 (Costa-Jussà et al., 2022) — combined with back-translation (Sennrich et al., 2016) or transfer from related higher-resource Indic languages (see Ranathunga et al. 2023 for a survey).

Our submission takes a different route. The translation budget goes entirely to Gemini API calls, with BM25 handling example selection at inference time. This follows work on retrieval-augmented in-context translation (Vilar et al., 2023; Agrawal et al., 2023) showing LLMs can match or exceed fine-tuned baselines given well-chosen parallel demonstrations (Brown et al., 2020). Related work from our lab on multimodal low-resource data generation (Xiao et al., 2025) and cross-script parallel data (Driggers-Ellis et al., 2025) also informs the data preparation.

3 System Description

The overview of the pipeline is given in Figure 1. In this section, we will describe the phases of this pipeline from start to finish.

3.1 Data Preparation

The retrieval bank $\mathcal{R}$ is built from WMT26 training data and augmented with WMT23 pairs and publicly available parallel corpora. We retain 21 Manipuri and 3,554 Mizo pairs unique to WMT23 (Pal et al., 2023).

Category 1 Languages For Assamese, we add 142,080 pairs from Samanantar (Ramesh et al., 2022), 1,359 from Endangered Recipes Translated

500 (Karya, 2026, 2025; Joshi et al., 2025), and 2,009 from IndicGenBench FLORES (Singh et al., 2024) to the 54,000 WMT26 pairs. Khasi receives 1,048 Endangered Recipes pairs in addition to 26,000 WMT26 pairs, while Manipuri receives 2,009 FLORES pairs and 21 WMT23 pairs in addition to 23,687 WMT26 pairs.

We add no WMT-external data for the remaining Category 1 languages. Meitei-Mayek has 17,001 WMT26 pairs. Nyishi has 60,000 WMT26 pairs, and Mizo has 53,543 pairs across WMT26 and WMT23—49,989 and 3,554 pairs, respectively. We focus our data-collection efforts on the more resource-constrained Category 2 languages, although additional parallel data may exist for these three languages. Remaining paragraphs in this subsection will detail data augmentation for each Category 2 language separately.

Bodo We combine 15,214 WMT26 pairs with 134,931 pairs from the English–Bodo Parallel Dataset (Panjiyar), 1,497 Endangered Recipes pairs, and 2,009 FLORES pairs to produce the largest corpus of WMT and external data.

Karbi We add 382 pairs mined from *A Grammar of Karbi* (Konnerth, 2014) and 1,144 from *Karbi Texts* (Konnerth and Tisso, 2018) to the 1,000 WMT26 pairs.

Kokborok, Nagamese, and Tagin The corresponding WMT26 sets contain 2,269, 1,998, and 5,344 pairs. We add 6,078 Kokborok pairs from the Kokborok Digitalisation Project (Caswell et al., 2025; Jones et al., 2023), 8,771 English–Nagamese pairs from FNPC-25 (Maheshwari and Keshri, 2025), and 854 English–Tagin pairs from GinLish v0.1.0 (Dugi, 2024).

All sources are merged using SHA-256 source-side deduplication, Unicode normalization (NFC for Bengali and Devanagari and NFKC for Latin

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

Table 1: Sizes of the in-domain training data before and after augmentation. The primary system uses the **Train** data, while the contrastive system uses the **Pairs** data. The augmentation ratio is computed as the number of augmented sentence pairs divided by the number of base pairs. † **Pairs** recalculated from **Train**. They are equal because we use no external data for Mni_Mtei.

scripts), and length-ratio filtering. For noisier external sources, we optionally run LaBSE (Feng et al., 2022) cross-lingual similarity filtering.

3.2 In-Domain Data Filtering

We restrict retrieval to in-domain data because sentence pairs from unrelated domains have little lexical or semantic overlap with the WMT26 inputs. As a result, BM25 assigns them low relevance scores, making them unlikely to be selected as useful in-context examples. We use filtered, in-domain data in both our primary and contrastive systems. The final primary corpus sizes **Train** and number of **Pairs**, including contrastive data, appear in Table 1.

3.3 BM25 Retrieval

The source side of $\mathcal{R}$ is indexed with BM25 Okapi (Robertson and Zaragoza, 2009) ($k_1 = 1.5$, $b = 0.75$). Given a source sentence q , the system retrieves the top- r training pairs $\mathcal{R}_r(q) \subseteq \mathcal{R}$, skipping exact matches on the query and deduplicating by 60-character normalized prefix to prevent overlapping corpora (e.g., Samanantar vs. PMIndia) from flooding the prompt with near-identical examples.

Getting the index direction right matters. For $\text{En} \rightarrow \text{X}$, the index is built on the English sentences. For $\text{X} \rightarrow \text{En}$, the query is in the Indic language, so the index must be built on the Indic-language side. We had a bug in the AmericasNLP system (Dhawan et al., 2026b) where Spanish (Es) was always indexed regardless of direction, but it was overlooked because $\text{X} \rightarrow \text{Es}$ translation was not necessary. We

find and resolve the issue in this shared task.

Configurations additionally include d development exemplars $\mathcal{D}_d \subseteq \mathcal{D}$, selected by word-overlap ranking with leave-one-out exclusion. The prompt context becomes $P(q) = \mathcal{D}_d \cup \mathcal{R}_r(q)$.

3.4 Translation

Prompts go to Gemini 2.5 Flash (Comanici et al., 2025) via the google-genai SDK (temperature 0, thinking disabled, max output 512 tokens). We use a 2-key API pool with round-robin and quota failover to manage rate limits. Each of the eleven languages has its own system prompt encoding script, word order, and morphological type — the Khasi prompt warns about VOS order and diacritics, the Bodo prompt states that despite Devanagari script the language is Tibeto-Burman and not Hindi. We elaborate these prompting considerations in our appendix.

Post-processing strips model-generated prefixes, applies Unicode normalization, and for non-Latin scripts validates that output characters fall within the expected Unicode block. The validation is especially important for Meitei Mayek (U+ABC0–U+ABFF), where Gemini occasionally falls back to Bengali. A negligible fraction of output lines (4 of 21,599; 0.018%) are empty due to Gemini safety filters on specific political sentences that temperature-perturbed retry does not resolve.

3.5 Relation to the AmericasNLP System

As shown in Figure 1, the main architectural change from the AmericasNLP pipeline (Dhawan et al., 2026b) is the removal of the vision-captioning stage. We also add bidirectional translation support through the BM25 index fix described, and merge multiple external corpora into retrieval banks.

4 Languages

Most language tracks in the shared task—eight of eleven—represent Tibeto-Burman languages, while the three non-Tibeto-Burman languages are Assamese, Khasi, and Nagamese (Post and Burling, 2017). The languages are predominantly Subject–Object–Verb (SOV), consistent with the clause-final pattern common across the Tibeto-Burman and local Indo-Aryan languages of Northeast India. Khasi, an Austroasiatic language, is the principal exception, with basic Subject–Verb–Object (SVO) order (Post and Burling, 2017; Sharma, 1999). Assamese is an Indo-Aryan language, while

Nagamese is an Assamese-lexified creole spoken across Nagaland (Post and Burling, 2017; Maiti et al., 2025). Manipuri appears in two task tracks, one written in Bengali script and the other in Meitei Mayek (Singh, 2012). These tracks represent two scripts of the same language rather than genetically distinct languages; we refer to the Meitei Mayek track as Meitei elsewhere in the paper. Bodo belongs to the Tibeto-Burman family but is written in the Devanagari script (Post and Burling, 2017; Narzary et al., 2022). In our experiments, Gemini occasionally produced Hindi-like Bodo output, likely because Bodo shares the Devanagari script with Hindi and Nepali (Narzary et al., 2022). Explicitly identifying Bodo in the system prompt removed this behavior.

5 Experimental Setup

Experiments run on the University of Florida HiPer-Gator cluster. The translation pipeline is CPU-only (it makes API calls to Gemini) and runs on the `hpg-default` partition. The environment is a Python venv (not conda). GPU nodes (`hpg-b200`) and `mamba` (conda) environments are used only for NLLB-200 data generation. Evaluation uses SacreBLEU to measure ChrF (Popović, 2015) and ChrF++ (Popović, 2017).

Since no official WMT26 development set was released early on, we hold the last 500 pairs or fewer from each language’s **Train** data as the set of possible development exemplars $\mathcal{D}$ ($|\mathcal{D}| \leq 500$) before adding external data to obtain **Pairs**. We record $|\mathcal{D}|$ for each language in Table 2.

We conduct a pilot grid search over hyperparameters $r \in \{0, 20, 40, 80\}$ and $d \in \{0, 10, 20, 49\}$ for each language and direction — a 4×4 grid of 16 configurations per language–direction pair, evaluated on $n = 50$ held-out development sentences per cell. The pilot grid search evaluates each (r, d) cell on n sentences from this split and is complete across all 22 language–direction pairs. We submit the configuration that achieves the greatest ChrF score in the pilot grid search.

6 Results

In this section, we introduce the results of our method for translation of Category 1 and Category 2 languages in both the $\text{En} \rightarrow \text{X}$ and $\text{X} \rightarrow \text{En}$ translation tasks. Table 3 lists the performance metrics for our contrastive approach for each language pair and translation direction, evaluated with the augmented

Cat.	Language	ISO	$ \mathcal{D} $
1	Assamese	As	500
	Mizo	Lus	500
	Khasi	Kha	500
	Manipuri	Mni	500
	Meitei	Mni_Mtei	500
	Nyishi	Njz	500
2	Bodo	Bodo	500
	Kokborok	Trp	500
	Karbi	Mjw	96
	Nagamese	Nag	199
	Tagin	Tgj	226

Table 2: $|\mathcal{D}|$ for each language in the shared task. Note that $|\mathcal{D}| = 500$ except for some Cat. 2 languages.

development datasets prior to submission. We measure performance using the ChrF and ChrF++ metrics. The former determined what r, d pairs feature in our submissions, and we include ChrF++ for direct comparison of dev and test performance.

Table 4 lists contrastive performance for the languages and translation directions according to the BLEU and ChrF++ measurements published by the shared task organizers (Pakray et al., 2026). Other test-time metrics are available in the shared task findings paper (Pakray et al., 2026).

7 Ablations

In this section, we look at alternative approaches to the proposed method. We hope these ablations will guide future work by showing how to improve or not to improve on the current method and by deepening the discussion to follow about its strengths and weaknesses.

7.1 Multi30k Domain Mismatch

The clearest negative result concerns synthetic data augmentation, which we apply to positive effect in multiple previous MT experiments (Dhawan et al., 2026a,b).

We prepare but ultimately *exclude* synthetic data that would have originally augmented three of the eleven languages in the present task. Following the MultiScript30k methodology (Driggers-Ellis et al., 2025), we generate NLLB-200 translations of the Multi30k image captions (Costa-Jussà et al., 2022; Elliott et al., 2016) into Assamese, Mizo, and Manipuri ($\sim 29\text{K}$ pairs each). This augmentation is the single largest driver of Guaraní performance in our AmericasNLP system. Multi30k captions are somewhat domain-matched to the captioning task (Dhawan et al., 2026b) because they are im-

Invariant					dev En→X				dev X→En			
Cat.	Language	ISO	Aug.	Pairs	<i>r</i>	<i>d</i>	ChrF	ChrF++	<i>r</i>	<i>d</i>	ChrF	ChrF++
1	Assamese	As	3.3×	171,794	80	10	52.95	49.81	40	49	68.80	67.98
	Mizo	Lus	1.0×	50,620	80	20	53.40	51.90	40	20	63.53	62.53
	Khasi	Kha	0.9×	22,847	80	49	56.44	55.71	80	49	75.64	75.09
	Manipuri	Mni	1.0×	22,544	80	49	66.72	62.34	20	49	69.97	67.86
	Meitei	Mni_Mtei	1.0×	16,312	80	10	41.22	36.86	40	10	55.20	52.85
	Nyishi	Njz	1.0×	49,793	80	20	44.98	40.58	80	20	58.76	57.96
2	Bodo	Bodo	9.3×	114,410	80	49	76.37	72.14	20	10	79.62	78.38
	Kokborok	Trp	1.0×	2,102	80	49	54.69	51.61	80	49	86.90	86.26
	Karbi	Mjw	1.0×	899	40	49	49.19	44.84	40	49	62.71	60.82
	Nagamese	Nag	5.7×	10,169	80	20	55.70	52.59	80	20	66.59	64.69
	Tagin	Tgj	0.9×	4,421	80	10	46.02	43.33	40	20	47.05	45.77

Table 3: Development contrastive submission results for each combination of language and translation direction. Two optimal *r*, *d* pairs are selected for each language from our search grid, one per translation direction. **Augmentation (Aug.)** gives a rough multiplier on the original size of the development set that our contrastive data augmentation provides. **Pairs** is the total number of contrastive retrieval exemplars $|\mathcal{R}|$ after augmentation.

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
	Mizo	Lus	1.0×	50,620	80	20	21.60	50.69	40	20	33.28	63.36
	Khasi	Kha	0.9×	22,847	80	49	29.14	53.16	80	49	25.76	52.92
	Manipuri	Mni	1.0×	22,544	80	49	8.15	39.89	20	49	45.93	72.37
	Meitei	Mni_Mtei	1.0×	16,312	80	10	16.62	52.85	40	10	38.13	66.16
	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
	Kokborok	Trp	1.0×	2,102	80	49	6.02	33.70	80	49	16.67	41.58
	Karbi	Mjw	1.0×	899	40	49	5.74	33.85	40	49	—	—
	Nagamese	Nag	5.7×	10,169	80	20	16.62	54.14	80	20	19.48	44.66
	Tagin	Tgj	0.9×	4,421	80	10	3.31	24.13	40	20	4.01	17.73

Table 4: Test contrastive submission results for each combination language and translation direction (Pakray et al., 2026). Two optimal *r*, *d* pairs are selected for each language from our search grid, one per translation direction. **Augmentation Ratio (Aug.)** gives a rough multiplier on the original size of the development set that our contrastive data augmentation provides. **Pairs** is the total number of contrastive retrieval exemplars $|\mathcal{R}|$ after augmentation. Entries in **bold** indicate that our submission performs best in the shared task for that metric.

age captions, and the queries at the MT stage of our submission are machine-generated Spanish captions of images. Meanwhile, WMT26 test data are sports news and government proceedings — “PCB requests ICC review,” “India loses 0-8 in hockey”.

We conclude that domain match contributes more strongly to performance than sheer volume in BM25-based RAG and exclude MultiScript30k augmentation from all final retrieval banks.

7.2 Key Observations from the Grid Search

Script effect. While they are treated as entirely separate languages by our method and the shared task documentation, we have already acknowledged that Manipuri is represented twice in the present task. What we call Manipuri is Manipuri in Bengali script while the language called Meitei is

Manipuri written in the Meitei Mayek script. For the Bengali script, our submission achieves 62.34 dev ChrF++ in the En→Mni direction, versus 36.86 dev ChrF++ for En→Mni_Mtei translation. This trend repeats for translation in the X→En direction but performance drops less steeply from 67.86 to 52.85 ChrF++. This represents roughly 25- and 15-point gaps for the same spoken language. We observe a similar trend for X→En translation on test data that reverses during En→X translation.

We conjecture that Gemini’s pre-training exposure to Manipuri in Bengali script is likely far greater than Manipuri in Meitei Mayek (U+ABC0–U+ABFF), which appears much more scarcely in online text, but that cannot account for the inconsistencies we observe. In any case, we caution that choice of orthography, where applicable, impacts

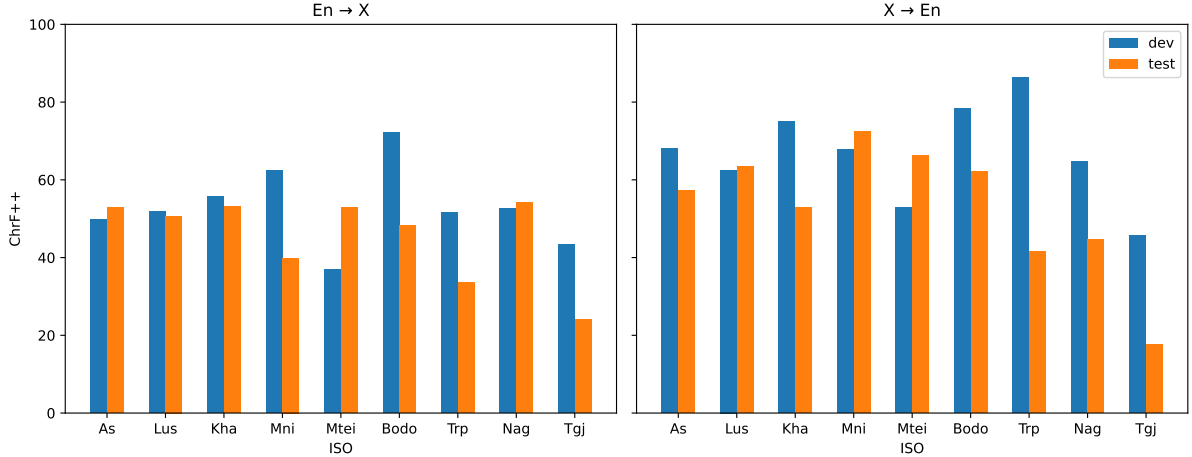


Figure 2: Dev-to-test performance changes in the contrastive system. **Left:** En→X translation performance for dev (blue) and test (orange) data. **Right:** X→En translation performance for dev (blue) and test (orange) data.

Invariant					En→X			X→En		
Cat.	Language	ISO	Aug.	Pairs	<i>r</i>	<i>d</i>	ChrF++	<i>r</i>	<i>d</i>	ChrF++
1	Assamese	As	3.3×	171,794	80	20	49.78	80	49	79.00
	Mizo	Lus	1.0×	50,620	80	49	52.54	40	49	77.90
	Khasi	Kha	0.9×	22,847	80	49	55.23	40	49	82.81
	Manipuri	Mni	1.0×	22,544	80	20	62.96	0	49	88.88
	Meitei	Mni_Mtei	1.0×	16,312	80	49	36.39	20	49	55.08
	Nyishi	Njz	1.0×	49,793	80	20	39.83	0	49	72.93
2	Bodo	Bodo	9.3×	114,410	80	20	73.36	0	49	81.87
	Kokborok	Trp	1.0×	2,102	80	49	52.76	0	49	68.29
	Karbi	Mjw	1.0×	899	40	49	46.76	40	49	62.71
	Nagamese	Nag	5.7×	10,169	40	10	53.87	0	20	53.52
	Tagin	Tgj	0.9×	4,421	80	49	35.31	0	49	71.16

Table 5: Results of repeating our r, d grid search on near-duplicate-filtered WMT26 data. **Augmentation (Aug.)** gives a rough multiplier on the original size of the development set that our contrastive data augmentation provides. **Pairs** is the total number of contrastive retrieval exemplars $|\mathcal{R}|$ after augmentation. Unchanged optimal r, d configurations in *italics*.

MT performance using our RAG method.

Data Cleaning. In addition to the original grid search described in Section 5, we perform the same procedure after the submission deadline using slightly different data and ranking. We filter the WMT26 training data again, this time removing near duplicate candidate exemplars, and rank on the ChrF++ metric rather than ChrF.

The results appear in Table 5, and comparison with Table 3 and Table 4 reveals that optimal the r, d configuration changes for 18 of 22 tasks. In particular, the optimal r falls to zero in seven out of eleven X→En translation tasks. Conversely, nine out of eleven X→En translation tasks now show $d = 49$. Six have ($r = 0, d = 49$). Nine out of eleven En→X tasks still show $r = 80$.

These results for newly filtered data, although

they still show performance for a small pilot of $n = 50$ samples, demonstrate a clear asymmetry with respect to translation direction that our original grid search does not show. An optimum $r = 80$ for most languages in En→X performance suggests that RAG does improve performance while generating a low-resource language, as in these tasks and our original AmericasNLP shared task submission. Two $r = 40$ configurations appear in Karbi and Nagamese translation, and these feature the smallest corpora, possibly indicating that retrieval strength is limited during extreme resource scarcity. Meanwhile, $r = 0$ in X→En tasks indicates that retrieval does not help when generating a high-resource language the LLM already knows well and when there are no samples identical or almost identical to the actual query to draw on.

Source-Side Retrieval Index. For $X \rightarrow \text{En}$, building the BM25 index on the Indic language’s side (rather than English in all settings) changes the optimal configuration for most languages. This is an essential improvement to our AmericasNLP 2026 submission (Dhawan et al., 2026b). Under the previous, source-invariant English BM25 index, the optimal r, d configuration had $r = 0$ for $\text{As} \rightarrow \text{En}$ and $\text{Kha} \rightarrow \text{En}$ translation. The implication is that no retrieval outperforms any retrieval based on the fixed English index for these languages. When the BM25 index is corrected, the optimal r increases to what we present in Table 3 and Table 4, indicating that retrieval does finally improve translation.

8 Discussion

For nearly every language and direction, we observe a some decay in performance from development to test set data. We expect this decay given our previous results in the AmericasNLP shared task (Dhawan et al., 2026b). The greatest dev-test differences appear for the Bodo and Kokborok and for $\text{En} \rightarrow \text{Bodo}$ translation and $\text{Trp} \rightarrow \text{En}$ translation in particular. We observe a difference of nearly 30 and 40 points, respectively, as we shift from the development to the test data. Still, Bodo and Kokborok are outliers as compared to the other languages, and there are many cases where test performance is actually greater than dev performance according to the ChrF++ metric. We build Figure 2 in Matplotlib (Hunter, 2007) to demonstrate the performance gaps and exceptions to predicted performance decline for both $\text{En} \rightarrow \text{X}$ (right) and $\text{X} \rightarrow \text{En}$ (left) translation tasks.

For languages where the gap is thinner or flipped, this may a signal that the test dataset for the current shared task is much closer to the same domain as the development set than the datasets in AmericasNLP 2026’s shared task, but a close examination of the test data or thorough documentation of it is necessary to confirm this hypothesis and it would not explain sizable performance gains as we move from development to test data.

Category 1 languages typically have many times more retrieval pairs than Category 2 languages, Bodo not withstanding. Least-squares linear regression analysis with Scipy (Virtanen et al., 2020), including Wald t-tests, reveals no significant ($p < 0.05$) correlation between **Pairs** and ChrF++ or BLEU performance, although we do uncover a weak but significant positive correlation ($r =$

0.650, $p < 0.05$, $m > 0$) between augmentation ratio and dev $\text{En} \rightarrow \text{X}$ performance. These findings may be linked to our previous observation that total retrieval bank volume is less important than domain alignment.

The effect of r and d values in this shared task is well explored through our grid search. For Category 2 languages, we observe that the gain from $d = 0$ to $d = 10$ exceeds the gain from $r = 0$ to $r = 80$, holding the other hyperparameter constant. Linear regression analysis also discovers a significant positive correlation between dev d values and $\text{En} \rightarrow \text{X}$ ChrF++ ($r = 0.625$, $p < 0.05$, $m > 0$) but none between r and $\text{En} \rightarrow \text{X}$ preformance. Curated development exemplars therefore appear to be a stronger signal than BM25-retrieved training pairs when the corpus is small. $r = 80$ performs best for ten of the eleven languages in $\text{En} \rightarrow \text{X}$ translation. In particular, Bodo performance peaks at ($r = 80, d = 49$) with a score of 76.37 ChrF and 72.14 ChrF++. We investigate the search grids further and find that the performance gains as r increases, holding d constant, are somewhat flat for languages with less than 10,000 total exemplars, including Karbi (899) and Tagin (4,421). Karbi is the one language with $r < 80$ in our contrastive $\text{En} \rightarrow \text{X}$ submissions. We conclude larger retrieval corpora better reward higher retrieval counts r .

9 Conclusion

We describe the gators submission to WMT26’s Indic Shared Task: A training-free, retrieval-augmented system that uses BM25 and Gemini 2.5 Flash to translate between English and eleven Northeastern Indian languages. The approach applies transfer learning to our AmericasNLP 2026 system (Dhawan et al., 2026b) but removes vision to isolate retrieval-augmented MT and builds retrieval banks with different parallel text.

Along with other ablations, we complete a grid search across all 22 language–direction pairs, and the submission covers every pair in both directions. We show the importance of domain matching to the synthetic data augmentation we espouse in the previous shared task (Dhawan et al., 2026b) and fix a critical bug in our BM-25 index construction.

While the previous shared task submission shows that retrieval-augmented MT can translate VLM-generated captions into low-resource languages, we now isolate the capabilities of retrieval-augmented MT and reproduce our previous suc-

Our submissions consistently outperform the competition by small yet measurable margins. Of 19 translation tasks, our contrastive submission achieves the best BLEU score for nine and the best ChrF++ score for eleven (Pakray et al., 2026).

Limitations

The system depends entirely on the Gemini 2.5 Flash API. During the pilot grid search, burst-level rate-limit 429 errors under exponential backoff consumed entire SLURM job windows, requiring careful scheduling and a 2-key API pool to complete all 22 directions’ grid cells in time for submission. A small number of test sentences (4 of 21,599; 0.018%) produced empty output due to Gemini’s safety filters blocking specific political content. In addition, we used (50) held-out development sentences for grid search due to Gemini API cost constraints. While a larger set could improve robustness, the selected configuration was sufficient to achieve competitive performance.

Retrieval bank quality varies widely. Category 2 languages have limited parallel data, and the available external corpora are often out-of-domain. For instance, our only external corpus for Nagamese consists of Bible translations. As an ablation in Section 7 shows, adding more data does not help when BM25 will not surface it.

The development split, consisting of the last 500 or fewer pairs of training data and with 50 sentences evaluated per grid cell, shares its domain with the training corpus; so dev metrics overstate performance. We see a similar effect in AmericasNLP (Dhawan et al., 2026b). In particular, Guaraní dev scores roughly halved during evaluation with the official test set. This time, Table 3 shows an inconsistent drop off for all languages except Bodo, but a performance drop usually exists.

Acknowledgments

We thank Arnold and Lisa Goldberg for their financial support and UF Research Computing for HiPerGator access. We also acknowledge the Gatorade, whose support provided the compute resources necessary for this research.

References

Sweta Agrawal, Chunting Zhou, Mike Lewis, Luke Zettlemoyer, and Marjan Ghazvininejad. 2023. In-context examples selection for machine translation.

In *Findings of the Association for Computational Linguistics: ACL 2023*.

Tom Brown et al. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901.

Minh Duc Bui, David Guzmán, Abteen Ebrahimi, Franklin Morales, et al. 2026. Findings of the AmericasNLP 2026 shared task on cultural image captioning for Indigenous languages. In *Proceedings of the Sixth Workshop on NLP for Indigenous Languages of the Americas (AmericasNLP)*, San Diego, California. Association for Computational Linguistics.

Isaac Caswell, Elizabeth Nielsen, Jiaming Luo, Colin Cherry, Geza Kovacs, Hadar Shemtov, Partha Talukdar, Dinesh Tewari, Baba Mamadi Diane, Koulako Moussa Doumbouya, Djibrila Diane, and Solo Farabado Cissé. 2025. [Smol: Professionally translated parallel data for 115 under-represented languages](#).

Gheorghe Comanici et al. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*.

Marta R Costa-Jussà, James Cross, Onur Çelebi, et al. 2022. No language left behind: Scaling human-centered machine translation. *arXiv preprint arXiv:2207.04672*.

Aashish Dhawan, Christopher Driggers-Ellis, Christan Grant, and Daisy Zhe Wang. 2026a. [Improving indigenous language machine translation with synthetic data and language-specific preprocessing](#). In *Proceedings of the Ninth Workshop on Technologies for Machine Translation of Low Resource Languages (LoResMT 2026)*, pages 119–126, Rabat, Morocco. Association for Computational Linguistics.

Aashish Dhawan, Christopher Driggers-Ellis, Dzmitry Kasinets, Christan Grant, and Daisy Zhe Wang. 2026b. 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.

Christopher Driggers-Ellis, Detraivious Brinkley, Ray Chen, Aashish Dhawan, Daisy Zhe Wang, and Christan Grant. 2025. [Multiscript30k: Leveraging multilingual embeddings to extend cross script parallel data](#).

Tungon Dugi. 2024. [GinLish v0.1.0: First-ever tagin and english parallel corpus](#). NIT Arunachal Pradesh. Dataset, version 0.1.0.

Desmond Elliott, Stella Frank, Khalil Sima’an, and Lucia Specia. 2016. [Multi30k: Multilingual english-german image descriptions](#). In *Proceedings of the 5th Workshop on Vision and Language*, pages 70–74. 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)*. Association for Computational Linguistics.
- Jay Gala, Pranjal A Chitale, et al. 2023. IndicTrans2: Towards high-quality and accessible machine translation models for all 22 scheduled Indian languages. *arXiv preprint arXiv:2305.16307*.
- J. D. Hunter. 2007. [Matplotlib: A 2d graphics environment](#). *Computing in Science & Engineering*, 9(3):90–95.
- Alex Jones, Isaac Caswell, Ishank Saxena, and Orhan Firat. 2023. [Bilex rx: Lexical data augmentation for massively multilingual machine translation](#).
- Neha Joshi, Pamir Gogoi, Aasim Mirza, Aayush Jansari, Aditya Yadavalli, Ayushi Pandey, Arunima Shukla, Deepthi Sudharsan, Kalika Bali, and Vivek Seshadri. 2025. [ELR-1000: A community-generated dataset for endangered indic indigenous languages](#). In *Proceedings of the 14th International Joint Conference on Natural Language Processing and the 4th Asia-Pacific Chapter of the Association for Computational Linguistics (IJCNLP-AACL 2025)*, Mumbai, India. Preprint: arXiv:2512.01077.
- Nabam Kakum, Sahinur Rahman Laskar, Koj Sambyo, and Partha Pakray. 2023. Neural machine translation for limited resources English–Nyishi pair. *Sādhanā*, 48(4).
- Karya. 2025. [ELR-1000: A community-generated dataset for endangered indic indigenous languages](#). Hugging Face.
- Karya. 2026. [Endangered recipes translated 500](#). Hugging Face.
- Linda Konnerth. 2014. *A Grammar of Karbi*. Ph.D. thesis, University of Oregon.
- Linda Konnerth and Sikari Tisso. 2018. [Karbi texts: A fully glossed corpus of different genres](#). *Himalayan Linguistics*, 17(2):117–472.
- Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Multilingual denoising pre-training for neural machine translation. *Transactions of the Association for Computational Linguistics*, 8:726–742.
- Anant Maheshwari and Priya Keshri. 2025. [FNPC-25: First nagamese–english parallel corpus](#). Kaggle. ABV-IIITM Gwalior; supervised by Anjali; 8,771 English–Nagamese parallel sentence pairs; MIT License; accessed July 17, 2026.
- Agniva Maiti, Manya Pandey, and Murari Mandal. 2025. [NagaNLP: Bootstrapping NLP for low-resource Nagamese creole with human-in-the-loop synthetic data](#).
- Sanjib Narzary, Maharaj Brahma, Mwnthai Narzary, Gwmsrang Muchahary, Pranav Kumar Singh, Apurbal Senapati, Sukumar Nandi, and Bidisha Som. 2022. [Generating monolingual dataset for low resource language Bodo from old books using Google Keep](#). In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 6563–6570, Marseille, France. European Language Resources Association.
- Partha Pakray, Reddi Mohana Krishna, Santanu Pal, Advaita Vetagiri, Sandeep Kumar Dash, Arnab Kumar Maji, Saralin A. Lyngdoh, Lenin Laitonjam, Anupam Jamatia, Koj Sambyo, Ajit Das, and Riyanka Manna. 2025. Findings of WMT 2025 shared task on low-resource Indic languages translation. In *Proceedings of the Tenth Conference on Machine Translation (WMT)*, pages 532–553, Suzhou, China. Association for Computational Linguistics.
- Partha Pakray, Santanu Pal, Advaita Vetagiri, Reddi Krishna, Arnab Kumar Maji, Sandeep Dash, Lenin Laitonjam, Sarah Lyngdoh, and Riyanka Manna. 2024. Findings of WMT 2024 shared task on low-resource Indic languages translation. In *Proceedings of the Ninth Conference on Machine Translation (WMT)*, pages 654–668. Association for Computational Linguistics.
- Partha Pakray, Santanu Pal, Advaita Vetagiri, Kshetrimayum Boynao Singh, Sandeep Kumar Dash, Arnab Kumar Maji, Saralin A. Lyngdoh, Lenin Laitonjam, Anupam Jamatia, Ajit Das, Sunita Warjri, Uzzal Sharma, Koj Sambyo, and Riyanka Manna. 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.
- Santanu Pal, Partha Pakray, Sahinur Rahman Laskar, Lenin Laitonjam, Vanlalmuansangi Khenglawt, Sunita Warjri, Pankaj Kundan Dadure, and Sandeep Kumar Dash. 2023. [Findings of the WMT 2023 shared task on low-resource Indic language translation](#). In *Proceedings of the Eighth Conference on Machine Translation (WMT)*, pages 682–694. Association for Computational Linguistics.
- Jigar Panjiyar. [English to bodo parallel dataset](#). Kaggle. Licensed under the GNU Free Documentation License 1.3.
- Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318.
- Maja Popović. 2015. chrF: character n-gram F-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*. Association for Computational Linguistics.
- Maja Popović. 2017. chrF++: words helping character n-grams. In *Proceedings of the Second Conference*

- on Machine Translation. Association for Computational Linguistics.
- Mark W. Post and Robbins Burling. 2017. [The tibeto-burman languages of northeast india](#). In *The Sino-Tibetan Languages*, 2nd ed., edited by Graham Thurgood and Randy J. LaPolla, Routledge, pp. 213–242.
- Gowtham Ramesh, Sumanth Doddapaneni, Aravinth Bheemaraj, Mayank Jobanputra, Raghavan AK, Ajitesh Sharma, Sujit Sahoo, Harshita Diddee, J Mahalakshmi, 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.
- Surangika Ranathunga, En-Shiun Annie Lee, Mariana Prifti Skenduli, Ravi Shekhar, Mehreen Alam, and Rana Kaur. 2023. Neural machine translation for low-resource languages: A survey. *ACM Computing Surveys*, 55(11):1–37.
- 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)*.
- Stephen Robertson and Hugo Zaragoza. 2009. The probabilistic relevance framework: BM25 and beyond. *Foundations and Trends in Information Retrieval*, 3(4):333–389.
- 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)*.
- Hanjabam Surmangol Sharma. 1999. [A comparison between Khasi and Manipuri word order](#). *Linguistics of the Tibeto-Burman Area*, 22(1):139–148.
- Harman Singh, Nitish Gupta, Shikhar Bharadwaj, Dinesh Tewari, and Partha Talukdar. 2024. [Indicgenbench: A multilingual benchmark to evaluate generation capabilities of llms on indic languages](#).
- Thoudam Doren Singh. 2012. [Bidirectional Bengali script and Meetei Mayek transliteration of web-based Manipuri news corpus](#). In *Proceedings of the 3rd Workshop on South and Southeast Asian Natural Language Processing*, pages 118–123, Mumbai, India. The COLING 2012 Organizing Committee.
- David Vilar, Markus Freitag, Colin Cherry, Jiaming Luo, Viresh Ratnakar, and George Foster. 2023. Prompting PaLM for translation: Assessing strategies and performance. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*.
- Pauli Virtanen, Ralf Gommers, Travis E. Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, Evgeni Burovski, Pearu Peterson, Warren Weckesser, Jonathan Bright, Stéfan J. van der Walt, Matthew Brett, Joshua Wilson, K. Jarrod Millman, Nikolay Mayorov, Andrew R. J. Nelson, Eric Jones, Robert Kern, Eric Larson, C J Carey, İlhan Polat, Yu Feng, Eric W. Moore, Jake VanderPlas, Denis Laxalde, Josef Perktold, Robert Cimrman, Ian Henriksen, E. A. Quintero, Charles R. Harris, Anne M. Archibald, Antônio H. Ribeiro, Fabian Pedregosa, Paul van Mulbregt, and SciPy 1.0 Contributors. 2020. [SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python](#). *Nature Methods*, 17:261–272.
- Bushi Xiao, Qian Shen, and Daisy Zhe Wang. 2025. [From text to multi-modal: Advancing low-resource-language translation through synthetic data generation and cross-modal alignments](#). In *Proceedings of the Eighth Workshop on Technologies for Machine Translation of Low-Resource Languages (LoResMT 2025)*, pages 24–35, Albuquerque, New Mexico, U.S.A. Association for Computational Linguistics.

Appendix

This final supplementary section is dedicated to communicating fine-grained details of our submissions that are not in focus in the main body of our paper. We summarize in Section 9.1 the prompt engineering that goes into calling Gemini 2.5 Flash during each experiment described and during testing for each language and translation direction. Section 9.2 discusses the r, d grid search we conduct over a pilot sample of the available WMT26 data to select retrieval hyperparameter configurations for our final submission.

9.1 Prompts

In Table 6 and Table 7, we describe the considerations taken at the prompting stage for each language and translation direction. In general, we conditioned Gemini 2.5 Flash to generate the correct script and encoding during $En \rightarrow X$ translation through our prompts. We also inform Gemini about morphological considerations including word order, tonality, and type of agglutination in the Indic target language where applicable. Because Gemini is asked to generate En during $X \rightarrow En$ translation, we find it much less necessary to precondition generation in the high-resource language through extensive prompting.

9.2 r, d Search Grid

Similar to our AmericasNLP 2026 shared task submission (Dhawan et al., 2026b), 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. We designate ChrF as the primary metric for the primary grid search, and our submission for each language and translation direction has the optimal r, d hyperparameter configuration we determine through this grid search. After the submission deadline, we repeat the procedure with deduplicated samples and rank according to the ChrF++ metric.

Inclusion in this appendix of a full search over the eleven shared task languages in both translation directions, a total of 22 translation tasks, would imply tabulating 22×16 or 352 grid search cells in the following pages. This is doubled to 704 cells if we include the second search’s results. Instead, we refer the reader to our Github page, where we host a [bash script](#) that can recreate our grid search experiments using WMT26 data.

Language		En→X		X→En	
Lang.	ISO	URL	Considerations	URL	Considerations
Assamese	As	URL	Agglutinative Postpositions Bengali-Assamese Script Honorifics Match Example Register Match Example Style No Addtl. En, Expl., Parens Output Only Single Line of As SOV Word Order Unicode NFC Encoding Unique Assamese Characters	URL	Agglutinative Bengali-Assamese Script Match Register of Examples No Addtl. Explanations Output Fluent En Output Natural En Output Only Single Line of En SOV Word Order
Mizo	Lus	URL	Latin Script Match Example Register Match Example Style No Addtl. Explanations Output Only Single Line of Lus Output Std. Lus Orthography SOV Word Order Tonal w/o Changes in Script Tibeto Burman Family Verb-Final Agglutination	URL	No Addtl. Explanations Output Fluent En Output Natural En Output Only Single Line of En
Khasi	Khasi	URL	Austroasiatic Family Different Typology Gendered Prefix Agreement Latin Diacritics No Addtl. Explanations Output Only Single Line of Kha Polysynthetic Morph. Preserve Diacritics VOS Word Order	URL	Gendered Prefix Agreement No Addtl. Explanations Output Fluent En Output Natural En Output Only Single Line of En Polysynthetic Morph. VOS Word Order
Manipuri	Mni	URL	Bengali Script Complex Verb Agglutination Match Example Style NFC Unicode Encoding No Addtl. Explanations No Bengali Output Only Single Line of Mni SOV Word Order Tibeto-Burman Family Unicode NFC Norm.	URL	No Addtl. Explanations Output Fluent En Output Natural En Output Only Single Line of En
Meitei	Mni_Mtei	URL	Agglutinative Match Example Script Match Example Style Meitei Mayek Script No Addtl. Explanations No Bengali Script No Devanagari Script Output Only Single Line of Mtei SOV Word Order Tibeto-Burman Family	URL	No Addtl. Explanations Output Fluent En Output Natural En Output Only Single Line of En
Nyishi	Njz	URL	Agglutinative Follow Example Orthography Latin Script No Addtl. Explanations Output Only Single Line of Njz Prefer Short Translations SOV Word Order Tibeto-Burman Family Very LoRes	URL	No Addtl. Explanations Output Fluent En Output Natural En Output Only Single Line of En

Table 6: We describe the prompting strategies we utilize for each Category 1 language in this table. Links in this table direct the reader to each prompt where it lives in our public Github repository.

Language		$En \rightarrow X$		$X \rightarrow En$	
Lang.	ISO	URL	Considerations	URL	Considerations
Bodo	Bodo	URL	Agglutinative Morph. Devanagari Script Examples Define Bodo Follow Examples Carefully Grammar Different From Hindi No Addtl. Explanations No Hindi, Sanskrit, or Marathi Not Indo-Aryan Family Output Only Single Line of Bodo SOV Word Order Tibeto-Burman Family Unicode NFC Normalization	URL	Devanagari Script No Addtl. Explanations Output Fluent En Output Natural En Output Only Single Line of En Tibeto-Burman Family
Kokborok	Trp	URL	Agglutinative Morph. Follow Examples Closely Latin Script Match Example Orthography Match Example Script No Addtl. Explanations Output Only Single Line of Trp SOV Word Order Tibeto-Burman Family Very LoRes	URL	No Addtl. Explanations Output Fluent En Output Natural En Output Only Single Line of En
Karbi	Mjw	URL	Agglutinative Extremely LoRes Latin Script Match Example Orthography No Addtl. Explanation Output Only Single Line of Mjw SOV Word Order Tibeto-Burman Family Tonal	URL	No Addtl. Explanations Output Fluent En Output Natural En Output Only Single Line of En
Nagamese	Nag	URL	En and As/Bengali Vocabulary En-Based Creole Fewer Inflections than En Srcs Latin Script Match Example Orthography No Addtl. Explanations No En No SOV Word Order Output Only Nag Creole Output Single Line of Nag Simplified Morphology Simplified En-Based Phonology SVO Word Order	URL	No Addtl. Explanations Normalize Nag Form to Std. En Output Fluent En Output Natural En Output Single Line of Std. En Output Std. En Only
Tagin	Tgj	URL	Agglutinative Extremely LoRes Latin Script Match Example Orthography Njz Relation No Addtl. Explanations Output Only Single Line of Tgj Shared Njz Vocab. SOV Word Order Tibeto-Burman Family	URL	No Addtl. Explanations Output Fluent En Output Natural En Output Only En Output Single Line of En

Table 7: We describe the prompting strategies we utilize for each Category 2 language in this table. Links in this table direct the reader to each prompt where it lives in our public Github repository.