

Winterpeg: Do Neural Cellular Automata Help Where Pretraining Ends?

Sujay Uday Rittikar¹, Sheela Ramanna¹

¹Applied Computer Science, The University of Winnipeg

Correspondence: s.ramanna@uwinnipeg.ca

Abstract

Machine translation for the low-resource languages of North-East India faces an uneven landscape: some are supported by large multilingual models, while others are absent from every pretrained system and usable subword vocabulary. For the covered pairs, fine-tuning a pretrained language model is a strong baseline, thus, our fine-tuned No Language Left Behind model (NLLB-200) ranks first in the WMT-2026 Indic MT shared task on English $\rightarrow$ Manipuri (Bengali-Assamese) at 11.95 BLEU and 44.67 chrF++. This paper addresses the remaining languages, for which no pretrained coverage exists. We train a vocabulary-free, UTF-8 byte-level, decoder-only language model from scratch and adapt it into a prompted translator that represents any script without tokenizer engineering. Operating at the byte level removes the vocabulary problem but shifts the cost onto the depth, as the model must compose byte sequences into the units, which a tokenizer would otherwise supply. We therefore introduce a causal Neural Cellular Automaton (NCA) as a front end, with a single local update rule iterated in place providing the required depth at the cost of a single set of shared weights. A front-end ablation at matched parameters attributes the observed gains to the iterated rule rather than to the addition of layers, while an equal-depth stack of independently parameterised layers recovers almost none of the improvement. Our causal NCA model places second on English $\rightarrow$ Mizo among primary submissions, degrades on Khasi, and reaches its limit on Meitei-Mayek, delineating a clear resource-and-script boundary from scratch byte-level modelling. Our code is available at https://github.com/sujayrittikar/wmt_2026_indic_task.

1 Introduction

Language embodies the collective memory and lived realities of its speakers. Machine translation is increasingly providing the essential bridge by

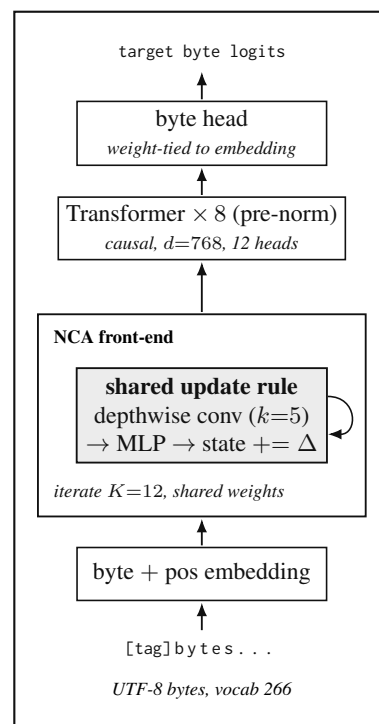


Figure 1: The byte-level Causal-NCA model. A weight-shared NCA update rule is iterated K times over the byte sequence before a causal Transformer and a weight-tied byte head.

allowing local knowledge to flow outward while inviting the global community to engage in return. For the languages of North-East India this bridge matters acutely. While Assamese, Manipuri, Mizo, and Khasi are spoken by millions, they coexist with English, which remains the dominant language for digital, administrative, and educational infrastructure. This creates a critical demand for translation to bridge the gap. The result is a familiar imbalance: abundant tools for a handful of high-resource languages and sparse support for low-resource languages. Building translation systems that serve these languages directly is therefore more than a technical exercise; it helps ensure their speakers are included as language technology advances.

Recent progress has concentrated almost entirely on large language models (Nyalang, 2026a; Team et al., 2022; Gala et al., 2023), for a good reason: their fluency and breadth are difficult to match. Yet, such a focus narrows the scope for the diverse methodologies necessitated by low-resource environments. A large model can only serve a language it has seen, and coverage follows speaker population and web presence, so the rarer the language, the less likely it is to appear in pretraining or to have a usable subword vocabulary. Those vocabularies come from tokenizers trained on corpora these languages barely populate, and which sometimes stereotype them (Nyalang, 2026b). Thus, relying solely on such dominant models fails to advance the languages that need the support the most.

In our submission to the WMT2026 low-resource Indic translation task (Pakray et al., 2026), we instead route each of six North-East Indian language pairs to the method suited to its resources. Where a language is well covered by pretraining, we fine-tune NLLB-200; where it is not, we translate with a vocabulary-free, byte-level model built on a causal NCA (Figure 1), trained from scratch on the little data these languages have. Bytes remove the vocabulary problem however, move the cost onto the depth, since the model must compose byte sequences into units a tokenizer would otherwise supply. We therefore front it with a causal neural cellular automaton (Mordvintsev et al., 2020), a single local update rule iterated K times, which supplies that depth at the cost of one set of weights. The sections that follow develop this routed design and examine how far the byte-level route carries the hardest, least-covered scripts in the corpora.

The rest of this paper is organized as follows. Section 2 reviews prior work on North-East Indian languages and low-resource machine translation. Section 3 describes the WMT2026 dataset. Section 4 presents the Causal-NCA model and our methodology. Section 5 details our experimental setup. Section 6 reports results, and Section 7 discusses them alongside other shared-task systems. Section 8 concludes, and Section 9 outlines limitations.

2 Related Works

Machine translation for North-East Indian languages has moved from statistical to neural methods, but parallel-data scarcity remains the central

obstacle. Early Manipuri work built bidirectional statistical systems from suffixes, dependency relations, and case markers (Singh and Bandyopadhyay, 2010) and compared the Bengali-derived and Meitei Mayek scripts (Singh, 2013), while Singh and Singh (2020) learned an unsupervised system from monolingual data alone. Subsequent efforts lean on two recurring levers, back-translation and explicit script or named-entity handling: back-translation for Mizo (Khenglawt et al., 2022) and Khasi (Nongbri et al., 2023), transfer learning for Khasi (Hujon et al., 2024), and a transliteration module for Assamese named entities and out-of-vocabulary tokens (Nath et al., 2025).

The WMT 2025 shared task (Pakray et al., 2025) set the current baselines, converging on parameter-efficient fine-tuning with back-translation and related-language data. TranssionMT (Zhou et al., 2025) led on Assamese, pairing IndicTrans2-1B (Gala et al., 2023) and NLLB-3.3B (Team et al., 2022) in a dual-model back-translation loop to reach 23.2 BLEU into English, with DoDS-IITPKD (Khongthaw et al., 2025) close behind, while English to Manipuri clustered near 4 BLEU across teams (Joshi et al., 2025; Acharya et al., 2025). Team ANVITA (J et al., 2025) selected their method by script support, extending NLLB-200’s vocabulary for Khasi, but falling back to a byte-level ByT5 model for the unsupported Nyishi script - a direction relevant to our methodology. Closest to our approach is TextNCA (Mittal et al., 2026), which uses a Neural Cellular Automaton as a standalone language model, whereas we employ it as a lightweight front end before a Transformer stack.

3 Dataset

We build our systems for six English-centric low-resource pairs covering several language families and three scripts (Table 1). Our training data combines three sources: (i) the parallel training corpora released by the WMT2026 shared task (Kakum et al., 2023; Pal et al., 2023; Pakray et al., 2024); (ii) the WMT2025 gold test sets, which we repurpose as additional parallel training data; and (iii) monolingual text used only for pretraining the byte-level model (Section 3.3).

3.1 Parallel data and cleaning

The organizer-released corpora has a total 230,687 sentence pairs. We apply a uniform cleaning

Pair (En-X)	ISO	Script	Family	Raw	Cleaned	+2025	Final
Assamese	asm	Eastern Nagari	Indo-Aryan	54,000	49,116	2,000	51,056
Manipuri	mni	Eastern Nagari	Tibeto-Burman	23,687	22,826	2,000	24,734
Meitei-Mayek	mni	Meitei Mayek	Tibeto-Burman	17,000	16,764	–	16,764
Mizo	lus	Latin	Kuki-Chin	50,000	49,144	2,000	51,043
Khasi	kha	Latin	Austroasiatic	26,000	25,836	2,000	27,783
Nyishi [†]	njz	Latin	Tani	60,000	50,046	2,000	51,961
Total				230,687	213,732	10,000	223,341

Table 1: Parallel training data by language. **Raw**: organizer release; **Cleaned**: after filtering/de-duplication; **+2025**: added WMT2025 gold pairs during final test-set evaluation; **Final**: merged, de-duplicated corpus. Manipuri and Meitei-Mayek are the same language (Meiteilon) in two scripts. [†]Nyishi was withdrawn from the shared task and excluded from official results.

pipeline: Unicode NFC normalization and whitespace squashing; exact and near-duplicate removal (casefolded, punctuation-stripped key); removal of empty and copy (src=tgt) pairs; and length filtering (max. 100 words / 1000 characters, word-ratio ≤ 3 , char-ratio ≤ 9). Cleaning removes 7.3% of pairs overall, but very unevenly: 16.5% for the dictionary-style Nyishi corpus (largely exact duplicates) versus roughly 1% for Meitei-Mayek. We additionally de-leak the corpus so that no source sentence in our held-out development set appears in training. To these we add the WMT2025 gold test sets (2,000 pairs each for Assamese, Khasi, Manipuri, Mizo, and Nyishi; no 2025 set exists for Meitei-Mayek). After merging and de-duplication, the final training corpus contains 223,341 sentence pairs (Table 1).

Pair	Median length	Tgt TTR	Char voc.	Scriptural %
Nyishi [†]	5	0.142	85	–
Manipuri	16	0.136	174	–
Meitei-Mayek	17	0.197	158	–
Assamese	19	0.087	182	–
Mizo	19	0.033	134	high
Khasi	28	0.008	82	~65

Table 2: Corpus characteristics by language. **Median length**: median English segment length in words; **Tgt TTR**: target-side type-token ratio; **Char voc.**: target character-vocabulary size; **Scriptural%**: share of segments containing a scriptural term.

3.2 Corpus characteristics

The pairs differ sharply in register and morphology (Table 2). Nyishi is short, dictionary-style text (median 5 English words per segment), whereas Khasi segments are longest (median 28). The Latin-script Bible-derived corpora are register-skewed: roughly 65% of Khasi and a comparable share of Mizo segments contain a scriptural term, which

Language	Wikipedia	GlottCC-V1	Sangraha	Total
Assamese	374,917	1,536	5,034,408	5,410,861
Manipuri	–	16,450	46,081	62,531
Meitei-Mayek	19,050	–	48,046	67,096
Mizo	–	99,167	–	99,167
Khasi	–	19,065	–	19,065
Nyishi [†]	–	23	–	23
Total	393,967	136,241	5,128,535	5,658,743

Table 3: Monolingual pretraining data (line counts after de-leaking). Wikipedia is CC BY-SA; GlottCC-V1 is CommonCrawl-derived; Sangraha is the AI4Bharat monolingual collection.

motivates a non-scripture held-out set. Target-side type-token ratios (TTR) confirm the skew: Khasi is extremely repetitive (TTR 0.008), while the Meitei-Mayek corpus is the most lexically diverse (TTR 0.197). Target character-vocabulary size relevant to our byte-level front-end ranges from 82 (Khasi, Latin) to 182 (Assamese, Eastern Nagari).

3.3 Monolingual data for pretraining

For byte-level pretraining we collect general-domain (deliberately non-scripture) monolingual text from Wikipedia (Wikimedia Foundation, 2023), GlottCC-V1 (Kargaran et al., 2024), and Sangraha (Khan et al., 2024), de-leaked against all development and test targets (Table 3). Coverage is highly uneven: Assamese has abundant text, whereas Nyishi is effectively absent (23 lines), reflecting its status as the lowest-resource language.

4 Methodology

We use two systems, routing each language pair to whichever performed better on our development sets (Table 4). Assamese and Manipuri, both well-covered by large multilingual models are handled by a fine-tuned NLLB-200. The remaining four pairs (Meitei-Mayek, Mizo, Khasi, and Nyishi)

are handled by a byte-level model with an NCA front-end that we train from scratch: Meitei Mayek, Khasi, and Nyishi are poorly covered by existing systems, while Mizo benefited more from our character-level model on the development set.

4.1 NLLB fine-tuning

For Assamese and Manipuri, we fine-tune nllb-200-distilled-600M, a single model serving both directions. Following FLORES-200 codes we use asm_Beng, mni_Beng, and eng_Latn. We train on all parallel data (including the WMT2025 gold pairs), both directions pooled, for 3 epochs (learning rate 3×10^{-5} , warmup, weight decay 0.01, fp16); as we retain no held-out split, the final epoch checkpoint is used. Decoding uses beam search (beam size 5).

Pair	System	Backbone
Assamese	NLLB	NLLB-200 (600M), fine-tuned
Manipuri	NLLB	NLLB-200 (600M), fine-tuned
Meitei-Mayek	NCA	Byte-LM (61M), from scratch
Mizo	NCA	Byte-LM (61M), from scratch
Khasi	NCA	Byte-LM (61M), from scratch
Nyishi [†]	NCA	Byte-LM (61M), from scratch

Table 4: System assignment per language pair. [†]Nyishi was withdrawn from the official results.

4.2 Byte-level Causal-NCA model

For the remaining pairs we train a single, vocabulary-free, decoder-only byte-level language model shared across all languages, then adapt it into a translator. Operating directly on UTF-8 bytes (vocabulary 266: 256 byte values, special tokens, and one tag per language) makes the model script-agnostic, sidestepping the poor subword vocabularies that tokenizers learn on tiny, unseen-script corpora.

Causal Neural Cellular Automaton. An NCA treats a signal as a grid of cells governed by a *single* learned update rule, applied repeatedly. Each cell refines its own state from a local view of its neighbours, and because the same rule runs at every position and every step, the front-end adds depth at a fixed parameter cost. We adapt this idea to autoregressive byte modelling, as illustrated in Figure 1. The byte sequence is a one-dimensional grid of cells: perception is a depthwise convolution (Gilpin, 2019) over neighbouring positions, and the update is a shared two-layer MLP returning a residual Δ added to the cell state. To keep

the model autoregressive we make the automaton *causal*: the perception convolution is left-looking (kernel 5, spanning positions $t-4$ to t), so no information flows from future bytes at any iteration, and a next-byte predictor never leaks context. Iterating this rule for K steps refines the byte representations without a fixed subword vocabulary, which is precisely what unseen scripts lack. We refer to this front-end as a *Causal-NCA*. The name denotes the weight-tied iterated block, and we do not rely on the self-organization or attractor behavior associated with image-domain NCAs. Section 6.3 demonstrates the iteration behavior.

Model. A byte and position embedding feeds the Causal-NCA front-end, followed by a standard pre-norm causal Transformer (8 layers, $d = 768$, 12 heads) and a weight-tied byte output head (61M parameters total). The front-end iterates its shared update rule $K = 12$ steps; because those steps reuse one set of weights, it adds representational depth at negligible parameter cost. During training, updates are stochastic (a random half of positions update per step, fire rate 0.5), and at inference, every position is updated, a difference Section 6.3 examines.

Pretraining. We pretrain the model as a byte-level language model on all six languages plus English, using both the parallel corpora and the monolingual data (Section 3.3), for 40k steps (AdamW, cosine schedule, peak learning rate 3×10^{-4}). We monitor per-language bits-per-byte.¹

Translation tuning. We adapt the pretrained model into a translator by prompted continuation: an input is encoded as [src-lang] source-bytes [tgt-lang] target-bytes [EOS], with the loss applied only over the target span. The source span attends bidirectionally (prefix-LM masking) while the target remains causal. We train one model per direction (mixing $\text{en} \rightarrow \text{X}$ and $\text{X} \rightarrow \text{en}$ in a single model degraded $\text{en} \rightarrow \text{X}$ quality), for up to 60k steps (learning rate 2×10^{-4} , cosine decay), alongside greedy decoding.²

4.3 Evaluation

We report chrF++, BLEU, TER (Translation Edit Rate), COMET, and BERTScore, treating chrF++

¹Pretrained checkpoint: <https://huggingface.co/sujayrattikar/ca-byte-lm-indic6>.

²Translation models: <https://huggingface.co/sujayrattikar/ca-byte-mt-en2x> ($\text{en} \rightarrow \text{X}$) and <https://huggingface.co/sujayrattikar/ca-byte-mt-x2en> ($\text{X} \rightarrow \text{en}$).

as our primary metric, as it suits morphologically rich and Latin-script targets on small test sets better than BLEU. We rank by the surface metrics (chrF++, BLEU, TER); TER matters here because it penalizes the length and repetition errors a from-scratch byte decoder is prone to, exposing degenerate output that overlap-based scores can mask. We treat COMET and BERTScore with more caution, as these embedding-based metrics are hard to interpret for scripts their encoders cover only sparsely, a point we revisit in Section 7.

chrF++ extends the character-level F-score of chrF (Popović, 2015) with word-level n-gram precision and recall. For a candidate c and reference r , the character n-gram precision and recall at order n are

$$P_{\text{ch},n} = \frac{\sum_{g \in \mathcal{C}_n(c)} \min(\text{cnt}_c(g), \text{cnt}_r(g))}{\sum_{g \in \mathcal{C}_n(c)} \text{cnt}_c(g)}, \quad (1)$$

$$R_{\text{ch},n} = \frac{\sum_{g \in \mathcal{C}_n(r)} \min(\text{cnt}_c(g), \text{cnt}_r(g))}{\sum_{g \in \mathcal{C}_n(r)} \text{cnt}_r(g)}. \quad (2)$$

where $\mathcal{C}_n(\cdot)$ is the multiset of character n-grams of order n and $\text{cnt}_x(g)$ counts occurrences of g in x . The per-order F_β -score with recall weight β is

$$F_{\beta,n} = \frac{(1 + \beta^2) P_n R_n}{\beta^2 P_n + R_n}, \quad (3)$$

applied to both character and word n-grams (default $\beta = 2$). Word-level precision $P_{w,n}$, recall $R_{w,n}$, and F-score $F_{w,n}$ are defined analogously over word n-grams. chrF++ averages both families:

$$\text{chrF++} = \frac{1}{N_{\text{ch}} + N_{\text{w}}} \left(\sum_{n=1}^{N_{\text{ch}}} F_{\text{ch},n} + \sum_{n=1}^{N_{\text{w}}} F_{\text{w},n} \right), \quad (4)$$

with default orders $N_{\text{ch}} = 6$ and $N_{\text{w}} = 2$.

TER measures the minimum number of edits needed to transform the candidate into the reference, normalized by reference length (Snover et al., 2006):

$$\text{TER} = \frac{I + D + S + M}{|\bar{r}|}, \quad (5)$$

where I , D , S , and M are the counts of insertions, deletions, substitutions, and phrase-level shifts (block movements) in the minimum-cost edit sequence, and $|\bar{r}|$ is the average reference length over all reference translations.

COMET is a learned, reference-based metric. A multilingual pre-trained encoder (XLM-R) maps the source s , hypothesis h , and reference r to pooled sentence embeddings $\mathbf{e}_s$, $\mathbf{e}_h$, $\mathbf{e}_r$, and a feed-forward estimator regresses these to human direct-assessment (DA) scores (Rei et al., 2020). The estimator input combines the embeddings with explicit cross-modal similarity features:

$$\mathbf{f} = [\mathbf{e}_s; \mathbf{e}_h; \mathbf{e}_r; \|\mathbf{e}_h - \mathbf{e}_r\|; \|\mathbf{e}_h - \mathbf{e}_s\|; \cos(\mathbf{e}_h, \mathbf{e}_r); \cos(\mathbf{e}_h, \mathbf{e}_s)], \quad (6)$$

$\text{COMET} = \text{MLP}(\mathbf{f})$.

where $\|\cdot\|$ is the Euclidean norm and $\cos(\cdot, \cdot)$ denotes cosine similarity.

5 Experimental Setup

All experiments were run on a single workstation with an AMD Ryzen Threadripper 5975WX CPU and 3 NVIDIA RTX A6000 GPUs (49 GB VRAM each), with training distributed across all GPUs.

6 Results

Table 5 reports our official WMT2026 primary submissions on the blind test sets, with the best system per direction for context. As primary metric we use chrF++; we also report BLEU and COMET, and give our rank among the primary submissions for each direction. Our two systems behave very differently, so we discuss them separately.

6.1 NLLB fine-tuning (Assamese, Manipuri-Bengali)

Fine-tuned NLLB-200 is our strongest system. On English→Manipuri (Bengali) it ranks **first** of 14 primary submissions on both chrF++ (44.67) and BLEU (11.95); in the reverse direction it places 3rd (65.55 chrF++, 37.81 BLEU). On Assamese it is competitive but middle-tier in both directions (6th; 48.32 / 51.34 chrF++).

6.2 Byte-level Causal-NCA (Mizo, Khasi, Meitei-Mayek)

The from-scratch NCA model demonstrates a sharp resource cliff. On **Mizo**, the largest of its pairs (49k) and Latin-script, it is competitive with fine-tuned multilingual systems: 2nd of 9 by BLEU on English→Mizo (19.66) and 4th of 8 on Mizo→English (58.96 chrF++), with the lowest

Direction	chrF++	Best chrF++	BLEU	TER↓	BERTScore	COMET	Rank
<i>NLLB-200 (fine-tuned)</i>							
En→Assamese	48.32	52.34	15.35	65.77	85.64	80.96	6/15
Assamese→En	51.34	57.43	21.57	64.66	94.29	84.85	9/13
En→Manipuri (Beng.)	44.67	44.67	11.95	74.93	87.12	70.36	1/14
Manipuri (Beng.)→En	65.55	72.44	37.81	44.56	95.77	82.68	3/14
<i>Byte-level Causal-NCA (from scratch)</i>							
En→Mizo	47.74	50.49	19.66	66.41	81.06	63.86	2/9
Mizo→En	58.96	63.36	30.53	52.32	95.29	79.34	4/8
En→Khasi	30.62	53.01	7.40	84.59	73.60	61.71	7/8
Khasi→En	18.14	53.15	0.51	112.93	84.22	40.72	8/9
En→Manipuri (Mtei.)	12.49	52.65	0.00	101.91	94.16	77.39	9/9
Manipuri (Mtei.)→En	17.33	66.38	0.00	122.02	83.49	40.19	8/9

Table 5: Official WMT2026 test results (Winterpeg primary submissions). **chrF++**: our primary system; **Best chrF++**: the top primary system in that direction; **TER↓**: lower is better; **Rank**: our position out of the number of primary submissions for that direction. Manipuri appears in both Bengali (Beng.) and Meitei Mayek (Mtei.) scripts.

k	LM	en→mizo	mizo→en
2	1.894	1.475	1.466
4	1.637	0.931	0.962
6	1.685	1.043	1.027
8	1.616	0.918	0.939
10	1.642	0.964	0.969
12	1.609	0.899	0.931
24	2.388	2.327	1.981

Table 6: Front-end truncation on Mizo development data. Bits-per-byte with the front-end stopped at k steps and the transformer untouched. All three checkpoints bottom out at $k = 12$. Lower is better.

TER (52.3) of any primary system in that direction. On **Khasi** it degrades but remains middle-tier on chrF++ for English→Khasi (30.62, 5th of 8), while Khasi→English is weak (18.14 chrF++, 0.51 BLEU). On **Meitei-Mayek** it fails outright, scoring 0 BLEU in both directions and placing last. Notably, this is not explained by language-model fit: the NCA achieved a low pretraining bits-per-byte on Meitei-Mayek (0.90, second only to Assamese/Manipuri), yet could not be tuned into a working translator, whereas Mizo, with a much higher bits-per-byte (1.68), translated well.

6.3 Front-end analysis for Mizo

We probe the contribution of the iterated front-end on Mizo checkpoints, our strongest byte-route pair, using held-out development text at inference only, with no retraining. As shown in Figure 2, the rule does not converge to a fixed point: (a) The residual norm $\|\Delta_k\|$ falls from 15.8 to 3.2 by $k=12$ but flattens onto a non-zero floor near 2.8. (b) The step-to-step contraction ratio rising toward one, so the decay is sub-geometric rather than contractive.

(c) The state norm grows roughly linearly, while similarity to the input keeps falling. The front-end therefore performs directional refinement with a growing norm, not relaxation to an attractor, and the iteration is thus better read as depth than as convergence to a self-organized state. The bits-per-byte bottoms out at $k \approx 12$ for the pretrained model and both translation directions, degrading monotonically beyond it (Table 6). The cause is a firing-rate mismatch, as each cell updates with probability 0.5 during training but 1.0 at inference with no expectation rescaling. Reproducing the training-time firing statistics at inference recovers most of the difference from full firing, and truncating to $k=12$ lands at the same floor (0.899).

6.4 Comparative Analysis of Direction pooling on Nyishi

Although Nyishi was withdrawn from the shared task this year, we use it here for corpus characterization and as a basis to compare direction pooling. Its scores were observed far above every other pair in this work, with 46.41 BLEU and 58.41 chrF++ on English→Nyishi, measured on the WMT-2025 test set (for this experiment, we did not add it in the training set). However, this reflects register rather than translation quality. The short formulaic segments may inflate all surface metrics. While we do not consider this evaluation comparable to the results of other languages, we use Nyishi to test a single model trained on both directions against one model per direction. Under pooling, the English→Nyishi dropped from 68.76 to 41.20 chrF++. A gap of this size suggests that the pooled model largely failed to learn the direction, rather than translating it less effectively. We observed

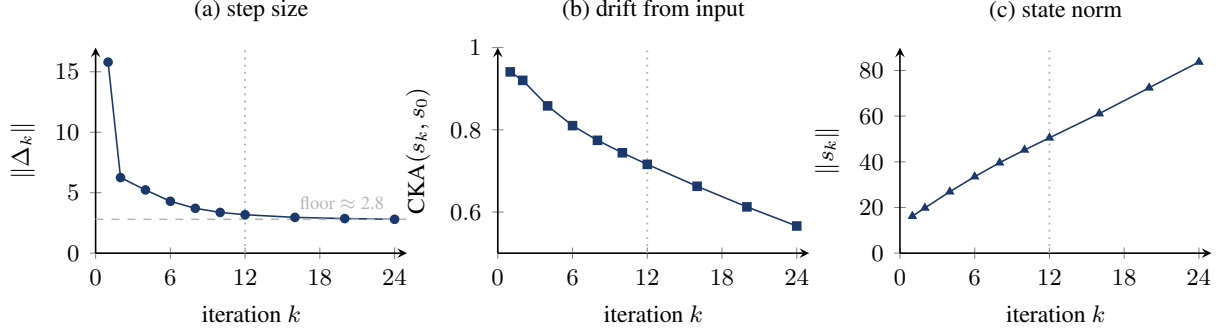


Figure 2: Front-end state dynamics on Mizo (single forward pass, no retraining). **(a)** The step size $\|\Delta_k\|$ settles onto a nonzero floor near 2.8 rather than vanishing. **(b)** Similarity to the input represented by Centered Kernel Alignment, $CKA(s_k, s_0)$, keeps falling through the trained horizon $K=12$. **(c)** The state norm grows roughly linearly. The rule therefore performs directional drift, not relaxation to a fixed point.

this behavior consistently across the other three languages as well. Although this is a single observation rather than a controlled study, it motivates our decision to train one model per direction throughout.

6.5 Ablation Study

We conduct ablation experiments to isolate the front-end as the only source of variation. The Table 7 presents a comparison between four front-end architectures under a fixed $d_{\text{model}}=768$ causal Transformer. The Causal-NCA arms use a single weight-tied update rule for $K \in \{1, 4, 12\}$ steps, and the untied convolutional stack applies 12 update layers of the same form, isolating the depth from weight sharing. The plain byte Transformer model, inspired from Xue et al. (2022), removes the front-end entirely and returns the freed 3.5M parameters to the feed-forward dimension, holding the comparison at a matched capacity. Finally, the *No front-end* architecture removes the front-end and runs at 57.3M parameters. The Plain byte Transformer and the model without any front-end perform equivalently on all three languages, differing by 0.3 chrF++ on Mizo, and only by 0.01 on Manipuri. However, allocating the same budget to a tied recurrent front-end at $K=12$ yields 42.53 chrF++ on Mizo against 34.43 for the baseline plain byte transformer, an increase of 8.10 chrF++. This pattern is consistent across all the metrics for all three languages. This improvement also follows the number of update steps rather than the number of parameterised layers. For instance, on Mizo, $K=1$ reaches 34.57 chrF++, and does not separate well from the baseline transformer. The untied stack, at an equal depth and matched parameter count, reaches 21.49

chrF++ on Manipuri (Mtei.), where it falls below the baseline Transformer.

7 Discussion

Where each system is suited. When a language is well covered by a large pretrained model, fine-tuning is difficult to improve upon. Our strongest result, English→Manipuri (Bengali), came from an NLLB model with 600M parameters and ranked first among primary submissions, ahead of several from-scratch and adapted systems. The vocabulary-free byte model with an NCA front-end is best suited to the pairs where such transfer is unavailable. On Mizo, a 61M from-scratch model comes within 1 to 3 chrF++ of fine-tuned multilingual systems and exceeds them on BLEU and TER, despite training on limited tokens. The two systems also differ in their decoding mechanism, since NLLB uses beam search while the byte-model decodes greedily. The causal NCA nonetheless records the lowest TER of any primary system on Mizo→English.

Where it breaks down. The byte model’s weakest results cluster on the two hardest settings, Meitei-Mayek (17k pairs, a unique abugida) and Khasi→English. Pretraining fit on Meitei-Mayek was good, so the shortfall lies in translation tuning rather than representation: with 17k parallel pairs and no monolingual English–Meitei signal, the greedy byte decoder does not learn to emit valid Meitei Mayek strings and drops to 0 BLEU. The recipe is therefore data-hungry at the tuning stage and brittle when the target script is both unique and low-resource. This underscores the need for caution regarding embedding-based metrics: Meitei-Mayek scores 0 BLEU yet 77.4 COMET and 94.2

Front-end	Params	English → X (held-out test)			
		chrF++	BLEU	TER↓	<i>p</i> vs. ctrl
<i>Mizo</i> , $n=2,000$ test					
Causal-NCA (tied), $K=12$	60.85M	42.53	18.52	69.36	0.001
Causal-NCA (tied), $K=4$	60.85M	39.47	17.64	70.14	0.001
Causal-NCA (tied), $K=1$	60.85M	34.57	14.05	77.46	0.072
Untied conv stack, $K=12$	60.83M	34.91	14.74	77.67	0.024
Plain byte Transformer (<i>control</i>)	60.84M	34.43	14.17	78.63	–
No front-end, d_{ff} unchanged	57.30M	34.73	14.56	78.35	0.070
<i>Khasi</i> , $n=1,293$ test					
Causal-NCA (tied), $K=12$	60.85M	35.79	13.43	78.88	0.001
Causal-NCA (tied), $K=4$	60.85M	34.52	12.66	80.89	0.001
Causal-NCA (tied), $K=1$	60.85M	33.69	13.03	82.73	0.001
Untied conv stack, $K=12$	60.83M	33.50	13.01	79.37	0.001
Plain byte Transformer (<i>control</i>)	60.84M	31.91	10.80	81.95	–
No front-end, d_{ff} unchanged	57.30M	32.62	8.97	82.35	0.064
<i>Manipuri</i> - Meitei-Mayek, $n=1,144$ test					
Causal-NCA (tied), $K=12$	60.85M	23.61	3.41	91.13	0.001
Causal-NCA (tied), $K=4$	60.85M	23.52	3.04	93.11	0.001
Causal-NCA (tied), $K=1$	60.85M	21.69	2.60	94.19	0.091
Untied conv stack, $K=12$	60.83M	21.49	2.82	96.37	0.160
Plain byte Transformer (<i>control</i>)	60.84M	21.63	3.11	94.26	–
No front-end, d_{ff} unchanged	57.30M	21.62	2.95	95.78	0.181

Table 7: Byte-level front-end ablation, English → X. All arms share an identical 8-layer, $d_{\text{model}}=768$ causal Transformer and differ only in the module placed between the byte embedding and the stack. Parameter counts are matched to within 0.03% (60.85 / 60.84 / 60.83M), except for *No front-end, d_{ff} unchanged*, which is 3.5M lighter at 57.30M. *p*-values are computed via paired bootstrap resampling (1000 samples) against the plain byte Transformer model as control for each language; bold indicates $p < 0.05$.

BERTScore, which their encoders cannot justify for a script they barely cover, so we treat the surface metrics (BLEU, chrF++, TER) as the trustworthy signal.

8 Conclusion and Future Work

We described Winterpeg’s WMT2026 Indic MT submission, a two-system approach: fine-tuned NLLB-200 for the well-covered pairs (Assamese, Bengali-script Manipuri) and a vocabulary-free byte model, built from the ground up and integrated with a Neural Cellular Automaton front-end for the rest. The byte-level Causal-NCA model places 2nd on English→Mizo BLEU among primary systems but degrades on Khasi and fails on Meitei-Mayek, tracing a clear resource-and-script cliff. Two key directions emerge: The *first* is to strengthen the byte model itself on the hardest pairs. A higher-capacity Causal-NCA (deeper or wider update rule, correcting the firing schedule) paired with better decoding than greedy search, such as beam search or sampling with repetition penalties, to curb the degenerate high-TER output we observe on Khasi and Meitei-Mayek. The *second* is to supply the target-side data these pairs lack, scal-

ing back-translation and monolingual pretraining and pooling Bengali- and Meitei-script Manipuri (the same language, Meiteilon) into a shared target space with transliteration-based augmentation.

9 Limitations

System assignment was chosen by development-set performance, not a controlled comparison, so we cannot isolate the NCA front-end’s contribution from data quantity or script difficulty; the front-end ablation remains future work. The byte model is trained on only a few million tokens with greedy decoding, so its scores are a lower bound. The submitted front-end also runs past its useful iteration count because of a mismatch between training and inference firing rates. We report automatic metrics only, and embedding-based metrics (COMET, BERTScore) are unreliable here, as shown by our 0-BLEU/77-COMET Meitei-Mayek result.

Acknowledgments

This research is supported by NSERC Discovery grant #194376. We thank the reviewers for their constructive feedback, and in particular Reviewer #1, whose detailed comments were highly

valuable to materially improve the clarity of this paper.

References

- Priyobroto Acharya, Haranath Mondal, Dipanjan Saha, Dipankar Das, and Sivaji Bandyopadhyay. 2025. [JU-NLP: Improving low-resource Indic translation system with efficient LoRA-based adaptation](#). In *Proceedings of the Tenth Conference on Machine Translation*, pages 1201–1209, Suzhou, China. Association for Computational Linguistics.
- Jay Gala, Pranjal A Chitale, A K Raghavan, Varun Gumma, Sumanth Doddapaneni, Aswanth Kumar M, Janki Atul Nawale, Anupama Sujatha, Ratish Pudupully, Vivek Raghavan, Pratyush Kumar, Mitesh M Khapra, Raj Dabre, and Anoop Kunchukuttan. 2023. [Indictrans2: Towards high-quality and accessible machine translation models for all 22 scheduled indian languages](#). *Transactions on Machine Learning Research*.
- William Gilpin. 2019. [Cellular automata as convolutional neural networks](#). *Physical Review E*, 100:032402.
- Aiusha Vellintihun Hujon, Thoudam Doren Singh, and Khwairakpam Amitab. 2024. [Neural machine translation systems for english to khasi: A case study of an austroasiatic language](#). *Expert Systems with Applications*, 238:121813.
- Sivabhavani J, Daneshwari Kankanwadi, Abhinav Mishra, and Biswajit Paul. 2025. [ANVITA : A multi-pronged approach for enhancing machine translation of extremely low-resource Indian languages](#). In *Proceedings of the Tenth Conference on Machine Translation*, pages 1240–1247, Suzhou, China. Association for Computational Linguistics.
- Nisheeth Joshi, Palak Arora, Anju Krishnia, Riya Lonchenpa, and Mhasilenuo Vizo. 2025. [BVSLP: Machine translation using linguistic embellishments for IndicMT shared task 2025](#). In *Proceedings of the Tenth Conference on Machine Translation*, pages 1265–1270, Suzhou, China. Association for Computational Linguistics.
- Nabam Kakum, Sahinur Rahman Laskar, Koj Sambyo, and Partha Pakray. 2023. [Neural machine translation for limited resources English-Nyishi pair](#). *Sādhana*, 48(4):237.
- Amir Hossein Kargaran, François Yvon, and Hinrich Schütze. 2024. [Glotcc: An open broad-coverage commoncrawl corpus and pipeline for minority languages](#). In *Advances in Neural Information Processing Systems*, volume 37, pages 16983–17005. Curran Associates, Inc.
- Mohammed Safi Ur Rahman Khan, Priyam Mehta, Ananth Sankar, Umashankar Kumaravelan, Sumanth Doddapaneni, Suriyaprasaad B, Varun G, Sparsh Jain, Anoop Kunchukuttan, Pratyush Kumar, Raj Dabre, and Mitesh M. Khapra. 2024. [IndicLLMSuite: A blueprint for creating pre-training and fine-tuning datasets for Indian languages](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15831–15879, Bangkok, Thailand. Association for Computational Linguistics.
- Vanlalmuansangi Khenglawt, Sahinur Rahman Laskar, Partha Pakray, Riyanka Manna, and Ajoy Kumar Khan. 2022. [Machine translation for low-resource english-mizo pair encountering tonal words](#). *Computación y Sistemas*, 26(3).
- Ontiwell Khongthaw, G.I. Salvin, Shrikant Budde, Abigail Chigwededza, Dhruvadeep Malkar, and Swapnil Hingmire. 2025. [DoDS-IITPKD:submissions to the WMT25 low-resource Indic language translation task](#). In *Proceedings of the Tenth Conference on Machine Translation*, pages 1248–1252, Suzhou, China. Association for Computational Linguistics.
- Avni Mittal, Avinash Anand, Ashutosh Kumar, Dikshant Kukreja, Kritarth Prasad, Sushane Dulloo, Erik Cambria, Timothy Liu, Zhengkui Wang, and Rajiv Ratn Shah. 2026. [Textnca: Neural cellular automata for language modeling via hierarchical local attention](#). *Preprint*, arXiv:2608.02050.
- Alexander Mordvintsev, Ettore Randazzo, Eyvind Niklasson, and Michael Levin. 2020. [Growing neural cellular automata](#). *Distill*. <https://distill.pub/2020/growing-ca>.
- Basab Nath, Sunita Sarkar, Somnath Mukhopadhyay, and Arindam Roy. 2025. [Improving neural machine translation by integrating transliteration for low-resource english–assamese language](#). *Natural Language Processing*, 31(2):306–327.
- Lavinia Nongbri, Gourashyam Moirangthem, Samarendra Salam, and Kishorjit Nongmeikapam. 2023. [Bidirectional neural machine translation \(NMT\) using monolingual data for Khasi-English pair](#). In *Proceedings of the 20th International Conference on Natural Language Processing (ICON)*, pages 318–325, Goa University, Goa, India. NLP Association of India (NLP AI).
- Badal Nyalang. 2026a. [NE-BERT: A multilingual language model for nine Northeast Indian languages](#). In *Proceedings of the Second Workshop on Language Models for Low-Resource Languages (LoResLM 2026)*, pages 1–12, Rabat, Morocco. Association for Computational Linguistics.
- Badal Nyalang. 2026b. [Stereotyped by silence: How LLMs erase Northeast Indian languages through omission and orthographic corruption](#). In *Proceedings of the 1st Workshop on Stereotypes Across Cultures in Language Technologies (StereACuLT 2026)*, pages 62–68, San Diego, California, United States. Association for Computational Linguistics.

- Partha Pakray, Reddi Krishna, Santanu Pal, Advaita Vetagiri, Sandeep 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*, pages 532–553, Suzhou, China. Association for Computational Linguistics.
- Partha Pakray, Santanu Pal, Advaita Vetagiri, Reddi Krishna, Arnab Kumar Maji, Sandeep Dash, Lenin Laitonjam, Lyngdoh Sarah, 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*, pages 654–668, Miami, Florida, USA. Association for Computational Linguistics.
- Partha Pakray, Santanu Pal, Advaita Vetagiri, Kshetri-mayum 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 2026 shared task on low-resource Indic languages translation. In *Proceedings of the Eleventh Conference on Machine Translation (WMT)*, Budapest, Hungary. Association for Computational Linguistics.
- 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*, pages 682–694, Singapore. Association for Computational Linguistics.
- Maja Popović. 2015. [chrF: character n-gram F-score for automatic MT evaluation](#). In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395, Lisbon, Portugal. Association for Computational Linguistics.
- Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. [COMET: A neural framework for MT evaluation](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2685–2702, Online. Association for Computational Linguistics.
- Salam Michael Singh and Thoudam Doren Singh. 2020. [Unsupervised neural machine translation for English and Manipuri](#). In *Proceedings of the 3rd Workshop on Technologies for MT of Low Resource Languages*, pages 69–78, Suzhou, China. Association for Computational Linguistics.
- Thoudam Doren Singh. 2013. [Taste of two different flavours: Which Manipuri script works better for English-Manipuri language pair SMT systems?](#) In *Proceedings of the Seventh Workshop on Syntax, Semantics and Structure in Statistical Translation*, pages 11–18, Atlanta, Georgia. Association for Computational Linguistics.
- Thoudam Doren Singh and Sivaji Bandyopadhyay. 2010. [Manipuri-English bidirectional statistical machine translation systems using morphology and dependency relations](#). In *Proceedings of the 4th Workshop on Syntax and Structure in Statistical Translation*, pages 83–91, Beijing, China. Coling 2010 Organizing Committee.
- Matthew Snover, Bonnie Dorr, Rich Schwartz, Linnea Micciulla, and John Makhoul. 2006. [A study of translation edit rate with targeted human annotation](#). In *Proceedings of the 7th Conference of the Association for Machine Translation in the Americas: Technical Papers*, pages 223–231, Cambridge, Massachusetts, USA. Association for Machine Translation in the Americas.
- 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.
- Wikimedia Foundation. 2023. Wikimedia wikipedia dataset. <https://huggingface.co/datasets/wikimedia/wikipedia>. Snapshot 20231101 (configs 20231101.as, 20231101.mni).
- Linting Xue, Aditya Barua, Noah Constant, Rami Al-Rfou, Sharan Narang, Mihir Kale, Adam Roberts, and Colin Raffel. 2022. [Byt5: Towards a token-free future with pre-trained byte-to-byte models](#). *Transactions of the Association for Computational Linguistics*, 10:291–306.
- Zebiao Zhou, Hui Li, Xiangxun Zhu, and Kangzhen Liu. 2025. [TranssionMT’s submission to the Indic MT shared task in WMT 2025](#). In *Proceedings of the Tenth Conference on Machine Translation*, pages 1271–1275, Suzhou, China. Association for Computational Linguistics.