TIP-RBG-AI: Overcoming Orthographic Discrepancies via Algorithmic Translinear Pipelines and Phylogenetic Script Mapping

Barathi Ganesh HB and **Michal Ptaszynski**
Kitami Institute of Technology
Kitami, Hokkaido, Japan
hbbg.jp@gmail.com

Meenakshi and **Jairam R**
RBG AI Research, RBG.AI
Coimbatore, Tamil Nadu, India
research@rbg.ai

Abstract

This paper details the submission of team TIP-RBG-AI for the WMT26 Indic-MT shared task. Addressing severe data sparsity, unstandardized orthographies, and script-level discrepancies across ten Indic languages, we introduce a robust, resource-agnostic cross-lingual framework. Rather than relying on standard downstream parameter fine-tuning, our methodology optimizes zero-shot inference topologies across three massively multilingual machine translation architectures: NLLB-200, MADLAD-400, and IndicTrans2. To systematically resolve vocabulary mismatch and tokenizer fragmentation, we implement a three-tiered preprocessing pipeline comprising baseline untuned direct inference for native scripts, translinear normalization via algorithmic back-transliteration for Romanized textual representations, and phylogenetic script mapping for undocumented, scriptless vernaculars. Empirical evaluations demonstrate that structurally aligning orthographic decoding environments and exploiting cross-lingual family networks substantially enriches semantic fidelity. Without executing a single downstream parameter update, our purely zero-shot framework secured multiple podium placements, including second-place finishes in the Bodo and Manipuri (Meitei Mayek) to English tracks, validating the potential of frontend representation alignment for moderate low-resource translation, while also exposing clear limits at the most extreme end of data scarcity. The code used for reproducing the experiments is publicly available at https://github.com/rbg-research/EMNLP-2026.

1 Introduction

The rapid deployment of massively multilingual machine translation (MMMT) models has transformed natural language processing, enabling cross-lingual knowledge transfer across hundreds of languages. Architectures such as NLLB-200, MADLAD-400, and IndicTrans2 demonstrate exceptional semantic capacity within well-represented linguistic bounds, but the WMT26 Indic-MT shared task exposes their critical vulnerabilities when extended to resource-constrained, geographically isolated dialects. The task spans ten linguistically diverse Indic languages, split into Category 1 (moderate digital footprints) and Category 2 (extreme low-resource dialects with profound data sparsity). Crucially, this failure does not stem from inadequate representational capacity, but from vocabulary mismatch and tokenization-level structural conflict caused by orthographic divergence and the lack of standardized digital scripts.

The core of this architectural bottleneck lies in profound orthographic discrepancies and the absence of standardized digital scripts. While languages like Assamese and Manipuri possess established native Unicode formats, the vast majority of the target dialects such as Mizo, Khasi, Karbi, and Tagin are either documented in arbitrary Romanized text or exist as entirely scriptless vernaculars. When these unstandardized Latin strings are fed into standard sub-word tokenizers trained on global corpora, the words are fractured into semantically meaningless sub-tokens. This fragmentation immediately destroys morphosyntactic cohesion within the embedding space, isolating these dialects from the rich, pre-trained parameters of the underlying foundation models. The conventional approach to mitigating this divide relies heavily on resource-intensive downstream fine-tuning. However, when parallel corpora shrink to just a few thousand sentences as is the case for several Category 2 languages standard parameter updates become highly susceptible to catastrophic forgetting and gradient domination by higher-resource tracks.

To resolve this limitation without relying on destructive fine-tuning, this work introduces a highly adaptable, resource-agnostic cross-lingual framework. We reframe the challenge of low-resource translation not as a model optimization

Proceedings of the Eleventh Conference on Machine Translation, pages 2135–2145
October 28-29, 2026 ©2026 Association for Computational Linguistics

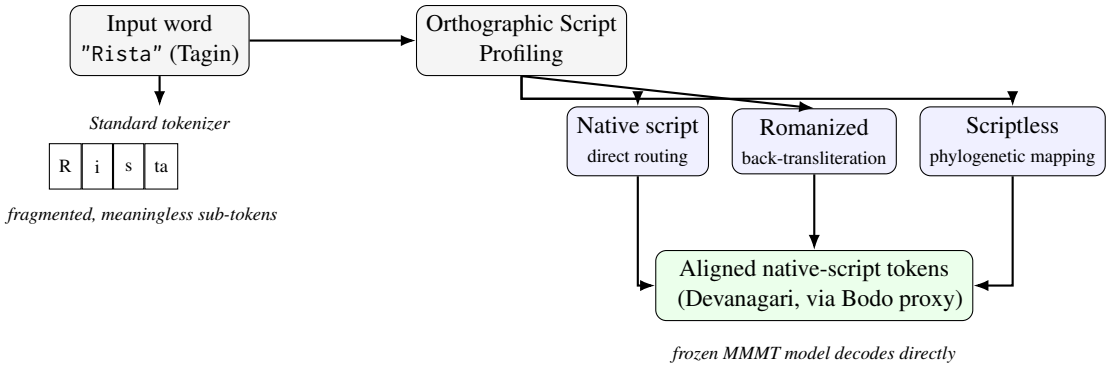


Figure 1: Overview of the problem and our approach. Standard tokenizers fragment Romanized and scriptless Indic words into meaningless sub-tokens (left), destroying cross-lingual alignment with pre-trained embeddings. Our Orthographic Script Profiling module routes each input through one of three tracks (native script, back-transliteration, or phylogenetic mapping) to restore native-script alignment before decoding, without any parameter updates to the frozen translation model.

problem, but as a frontend representation alignment issue. By optimizing zero-shot inference topologies across frozen foundation models, we demonstrate that latent translation capabilities can be unlocked purely through structural text normalization. Our methodology systematically addresses the data tiers defined by the organizers through a dynamic three-tiered preprocessing pipeline tailored to the exact orthographic profile of the input stream.

The primary scientific contributions of this work are realized through our pipeline customization. First, we bypass the fragmentation barrier for Romanized target inputs using an Algorithmic Translinear Pipeline, which systematically back-transliterates Latin strings into their regionally dominant Brahmic scripts to preserve morphosyntactic continuity. Second, for minority vernaculars completely lacking a digital script, we introduce a Phylogenetic Script Mapping heuristic. This novel routing paradigm leverages genetic language family alignments, syntactic typologies, and geographical continuums to anchor scriptless dialects to computationally stable proxy macro-languages. By strategically routing these normalized data streams through an automated policy engine to the most capable foundation model NLLB-200, MADLAD-400, or IndicTrans2, our empirical evaluations confirm that structured orthographic decoding substantially enriches semantic fidelity and lexical preservation for moderate low-resource task categories, while revealing that the most extreme low-resource dialects require parameter-level adaptation beyond what frontend alignment alone can provide.

The remainder of this paper is structured as follows. Section 2 reviews related work in multilingual translation and orthographic normalization. Section 3 details the extreme data imbalances and lexical disparities present in the WMT26 Indic-MT corpora. Section 4 outlines our translinear and phylogenetic methodology. Section 5 presents our zero-shot experimental design and model selection criteria, followed by a comprehensive performance analysis in Section 6. Finally, Section 7 concludes our findings and discusses future directions for resource-agnostic machine translation.

2 Related Work

The rapid evolution of massively multilingual machine translation (MMMT) has fundamentally expanded the limits of cross-lingual generation by scaling parameters and co-training on hundreds of languages simultaneously. Foundational architectures such as NLLB-200 (Costa-jussà et al., 2022) and MADLAD-400 (Kudugunta et al., 2023) utilize data-driven sampling strategies to build robust representation networks across shared semantic spaces. The creation of large parallel corpora in the Indian subcontinent, such as Samanantar (Ramesh et al., 2022), revealed the extreme data imbalance facing regional dialects. Such scaling of data eventually led to the development of translation models like IndicTrans2 (Gala et al., 2023), which optimizes tokenization arrays and pre-training methods to support Indic languages effectively. Yet, a survey of multilingual neural machine translation by Dabre et al. (2020) continuously stresses one crucial drawback: when such highly advanced models are applied zero-shot to extreme low-resource dialects, their semantic capacity across languages degrades

due to severe data imbalance and rapid vocabulary shifts.

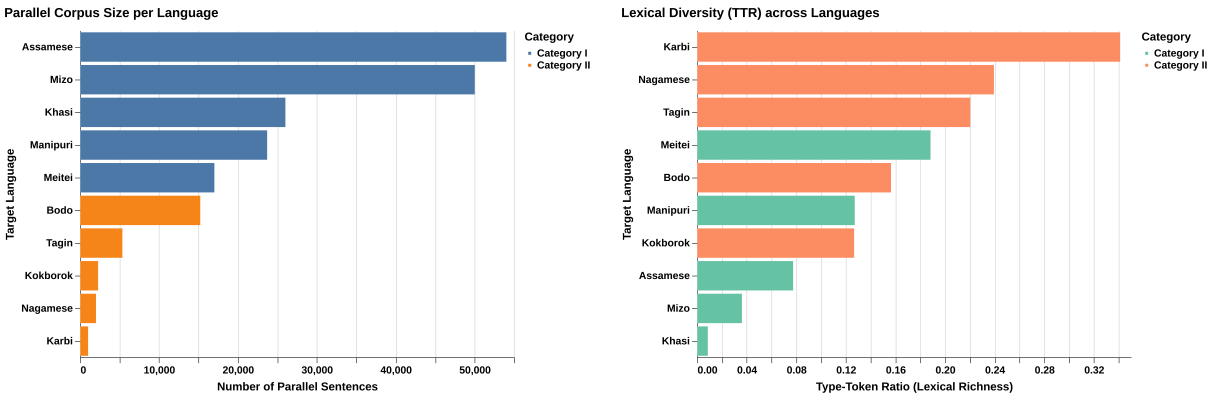
A primary cause of this degradation is the structural collapse of sub-word tokenization when challenged by script discrepancies and character-level noise (Salesky et al., 2020). In parallel, Rust et al. (2021) demonstrate that sub-word tokenizers trained on skewed, multilingual corpora frequently fragment underrepresented languages into semantically vacant sub-tokens, effectively destroying contextual meaning before the text reaches the transformer layers. This fragmentation is drastically compounded by the cross-script divide. As demonstrated by Libovický et al. (2022), multilingual language models naturally cluster their latent spaces according to script families. Consequently, feeding arbitrary Romanized text into an architecture optimized for Devanagari or Bengali scripts forces the model out of its learned distribution, leading to an explosion of unknown tokens (Pfeiffer et al., 2021). To circumvent these boundaries without fine-tuning, prior works have explored mapping phonological rules (Roark et al., 2020) and employing uniform input unification via romanization (Amrhein and Sennrich, 2022). Our methodology shows that the inverse approach, algorithmically reverting Latinized vectors back to native Brahmic script abugidas, restores tokenization stability and realigns inputs with the pre-trained embedding layers of frozen language models.

For entirely undocumented or scriptless vernaculars, translation efficacy traditionally drops to an absolute floor. To mitigate this, recent literature has pivoted toward exploiting the inherent typological similarities shared within language families to facilitate zero-shot knowledge transfer. Foundational work in neural machine translation established that models could achieve cross-lingual transfer by clustering linguistically related languages (Zoph et al., 2016; Nguyen and Chiang, 2017). Expanding on this, Lin et al. (2019) formulated systematic heuristics for selecting optimal transfer languages, proving that anchoring a low-resource dialect to a high-resource proxy based on phylogenetic lineage significantly outperforms generic multilingual transfer. Similarly, Oncevay et al. (2020) demonstrated that linking explicit linguistic typologies such as shared Subject-Object-Verb syntactic ordering, with architectural constraints helps stabilize translation pathways for resource-poor target systems. Our framework directly operationalizes these structural paradigms. By implementing a phylogenetic routing heuristic, we bypass the lack of native parameters for dialects like Tagin, Kokborok, and Karbi, dynamically mapping them into the computationally stable, genealogically aligned embedding spaces of their high-resource proxy anchors.

3 Data

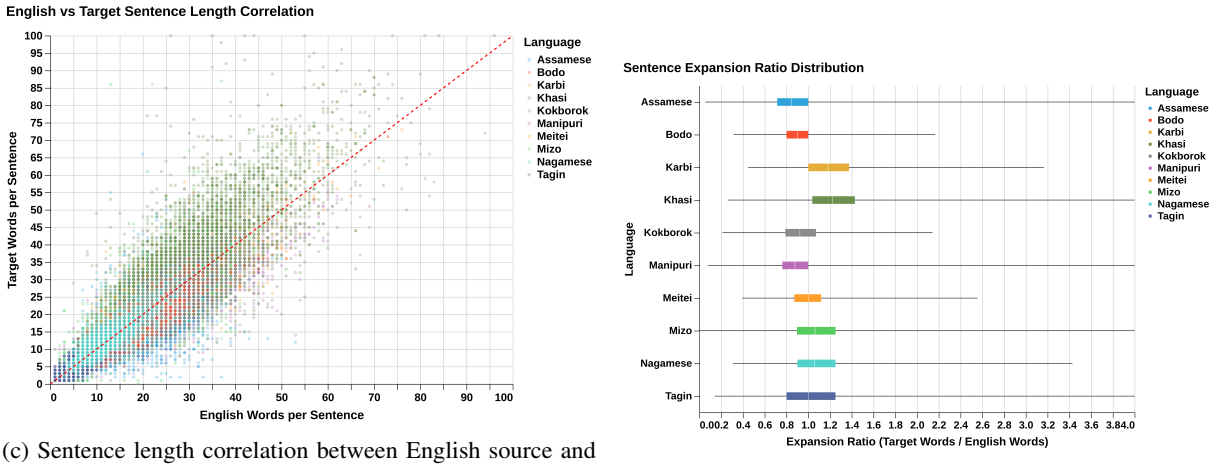
The parallel corpora provided for the WMT26 Indic-MT shared task present a highly asymmetric data landscape, characterized by steep resource imbalances and complex orthographic variations across ten distinct target languages (Pal et al., 2023; Pakray et al., 2025, 2024). The task organizers divided the dataset into two categories. Category 1 comprises languages with intermediate-level digital footprints: Assamese (54,000 sentences), Mizo (50,000), Khasi (26,000), Manipuri Bengali script (23,687), and Manipuri Meitei Mayek script (17,000). Category 2 comprises extremely low-resource languages with severe data scarcity: Bodo (15,215 sentences), Tagin (5,343), Kokborok (2,269), Nagamese (2,000), and Karbi (1,000). This drastic fall, from tens of thousands of parallel sentences to as few as 1,000, results in significant corpus imbalance.

The central architectural challenge inherent to this corpus stems from a profound cross-lingual script divide, which introduces structural friction during sub-word tokenization. The ten target languages are split into two mutually exclusive orthographic groups: Assamese, Bodo, and Manipuri (in both its Bengali and Meitei Mayek script forms) utilize traditional native scripts (employing complex Brahmic and indigenous character arrangements like Bengali-Assamese, Devanagari, and Meitei Mayek), whereas Mizo, Khasi, Nagamese, Karbi, Kokborok, and Tagin rely on Romanized orthographies using the Latin script. When attempting to leverage a single shared embedding space across a unified vocabulary, this structural split causes massive token fragmentation. Since character overlap between the native scripts and Romanized systems is virtually non-existent, standard byte-pair encoding algorithms trained uniformly on the raw text become heavily skewed. The larger Latin-based corpora easily dominate the vocabulary allocation, stripping native scripts of optimal sub-word representation. This structural complication forms the direct empirical justification for our architectural routing engine, which segregates native script



(a) Parallel corpus size (sentence count) across categories.

(b) Type-token ratio (TTR) across the target languages.



(c) Sentence length correlation between English source and target languages.

(d) Distribution of target-to-source word count expansion ratios.

Figure 2: Empirical analysis of the WMT26 Indic-MT shared task dataset. The plots highlight severe data imbalances (a), inflated lexical diversity due to orthographic non-standardization in Romanized tracks (b), and structural/geometric divergence between native-script compression and Latin-script expansion during translation (c, d).

tracks from romanized text streams before decoding.

However, the examination of type-token ratios in all these languages demonstrates a sharp difference between real morphological abundance and transliteration errors. Lexical diversity varies greatly within all the songs, showing an inverse correlation to corpus size. The low values of type-token ratios observed in Category 1 languages such as Khasi and Mizo indicate a very low degree of lexical diversity with the index of 0.0086 and 0.036, respectively. It shows extremely repetitive, dense structural token patterns, which means that although there is a large number of sentences in those languages, their vocabulary footprint is quite small. The opposite is seen in low-resource tracks, which have the highest type-token ratios: 0.156 for Bodo, 0.220 for Tagin, 0.239 for Nagamese, and the most surprising value of 0.341 for Karbi. In such Romanized environment, a sharp increase in vocabulary is not a sign of grammar complexity, but it is a sign of a serious amount of orthographic errors due to the absence of spelling rules in Latin alphabets. The translators often record the same phonemes using different characters, which leads to an increase in unique word tokens.

Geometric evaluations of sentence length correlations (Figure 2c) and expansion distributions (Figure 2d) further highlight the structural divergence separating the two language sets. When plotting the ratio of target words relative to the English source text, languages utilizing native scripts consistently demonstrate semantic contraction. Assamese displays a compression ratio of 0.87, followed closely by Manipuri at 0.91 and Meitei at 1.00. Indic abugidas inherently pack intricate phonetics, inflections, and postpositions into singular, compact word units, minimizing space-separated

markers. In contrast, Romanized tracks undergo massive morphosyntactic expansion, with expansion distributions reaching 1.09 for Mizo, 1.21 for Karbi, and peaking at 1.25 for Khasi. Mapping indigenous accents and phonemes into a restricted Latin character set forces the generation of multi-character digraphs and separate particle words, creating elongated point clouds that widen drastically as sentence lengths grow. Collectively, the data constraints, vocabulary skews, and geometric variations illustrated in Figure 2 establish the empirical foundation required to interpret the system performance discrepancies discussed in our subsequent analysis.

4 Methodology

To resolve the multi-dimensional challenges of severe data sparsity, unstandardized orthographies, and script-level discrepancies across the target languages without relying on resource-intensive downstream parameter fine-tuning, we introduce a robust, resource-agnostic cross-lingual framework. Our approach views orthographic discrepancies as a frontend representation alignment issue, transforming text at the boundary layer to fit the stable parameter space of frozen massively multilingual machine translation (MMMT) models. Formally, let X represent an input sentence in a target source language L_s, and let Y be the corresponding English sentence to be generated. Standard translation configurations parameterize the conditional probability $P(Y \mid X; \Theta)$ across an optimized model parameter set Θ. In sub-word tokenization architectures, arbitrary orthographic variations disrupt the sequence generation pipeline, yielding corrupted token arrays $\mathbf{t} = \text{Tokenizer}(X)$ that degrade cross-lingual transfer. We mitigate this structural decay by introducing a frontend transformation function $\Gamma(X)$, restructuring the string sequence so that the decoding target maps directly onto frozen language networks:

$$Y^* = \arg\max_{Y} P(Y \mid \Gamma(X); \Theta_{\text{frozen}}) \quad (1)$$

The execution of this transformation function is driven by an automated Orthographic Script Profiling module. In this module, the incoming sentence arrays will be intercepted and their characters' Unicode blocks scanned to dynamically classify the stream of texts into three processing paths according to script characteristics. The sentences classified under the Native Script processing path are texts written in standardized regional languages with native fonts, including Assamese written in Bengali script, Bodo written in Devanagari script, Manipuri written in Bengali script and Meitei written in Meitei Mayek script. These sentences will not be processed by any wrappers but will go through specialized model spaces for the sub-continent directly. Conversely, the Romanized track handles languages that present exclusively in Latin-transliterated text, such as Mizo, and Khasi. Direct tokenization of these strings usually fragments words into meaningless sub-tokens, destroying semantic cohesion. These sequences are diverted to an algorithmic translinear pipeline for script normalization. Finally, the Scriptless track isolates minority vernaculars or undocumented dialects that lack standardized digital orthographies entirely, including Karbi, Kokborok, Nagamese, and Tagin, redirecting them to a phylogenetic script mapping layer to anchor their representation space.

For sequences intercepted by the Romanized layer, the framework applies an Algorithmic Translinear Pipeline to convert Latin script representations back into native Brahmic structures. Let $\mathbf{w}_{\text{roman}} = [c_1, c_2, \ldots, c_n]$ represent a string character sequence of Latinized text. The translinear mapping layer transforms this input into an optimized native script vector $\mathbf{w}_{\text{native}}$ by maximizing the phonetic decoding probability across a generative beam:

$$\mathbf{w}_{\text{native}} = \arg\max_{\mathbf{w}} \prod_{i=1}^{m} P(c_i^{\text{native}} \mid c_{1:i-1}^{\text{native}}, \mathbf{w}_{\text{roman}}) \quad (2)$$

This operation is carried out using AI4Bharat's IndicXlit transliteration engine, a transformer-based back-transliteration model trained on the Dakshina and Aksharantar corpora for 21 Indic languages. We configure IndicXlit with a decoding beam width of 4 and apply phonetic rescoring by re-ranking the beam's n-best native-script candidates against a character-level language model trained on the target script's monolingual corpus, selecting the candidate that maximizes P(w_native) · P(w_native | w_roman). For languages without a direct IndicXlit model (e.g., Mizo, Khasi), we route through the closest supported script family (Bengali-Assamese) as the transliteration target before applying the phylogenetic proxy mapping described below

For undocumented or scriptless vernaculars that are completely omitted from the pre-training data

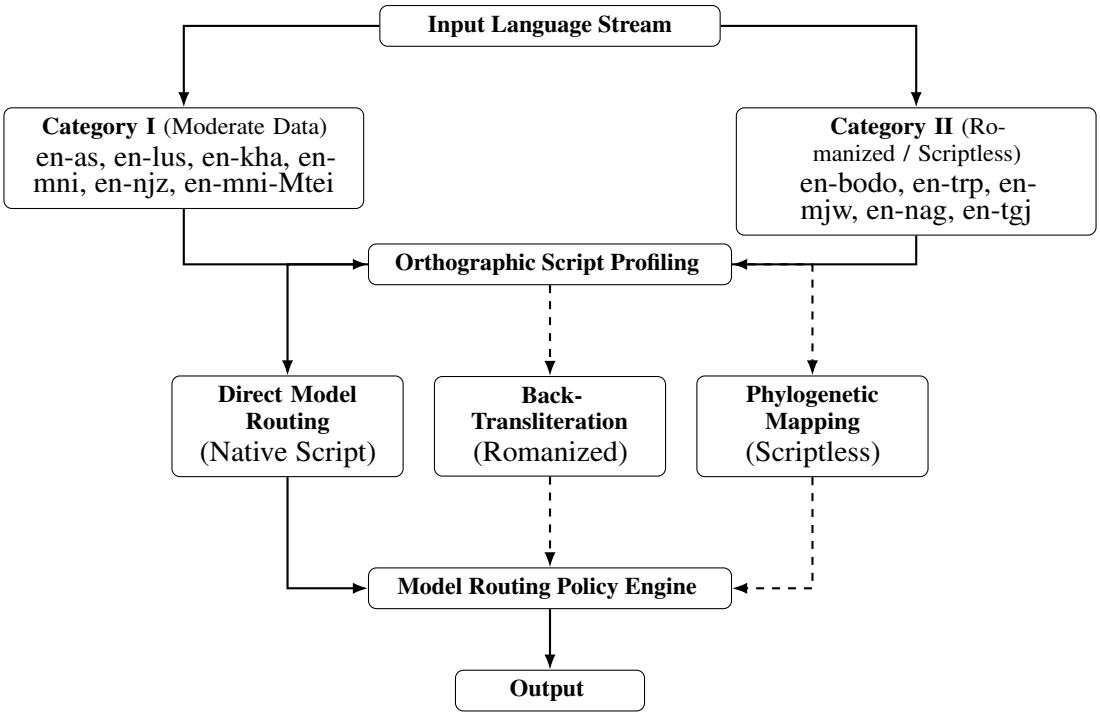


Figure 3: Architectural pipeline of the TIP-RBG-AI framework. Solid lines represent direct data flow paths where text requires no script conversion, whereas dashed lines represent transformed data flows that undergo frontend preprocessing (back-transliteration or phylogenetic mapping) to resolve orthographic discrepancies prior to final model routing.

of massive foundation architectures, translation typically drops to a baseline of zero. We resolve this bottleneck by introducing a Phylogenetic Script Mapping heuristic that establishes an operational representational space within the model's latent representation layers. This paradigm maps unrepresented target dialects onto well-supported Category 1 macro-languages by analyzing language family trees, morphological structures, word-order typologies (such as shared Subject-Object-Verb alignment), and regional geographical continuums. Through this genetic mapping process, Nagamese is routed to Assamese based on its characteristics as an Assamese-derived creole sharing core lexical inventories. Kokborok and Tagin are mapped directly to Bodo, leveraging their shared Tibeto-Burman lineage and regional continuums across Northeast India. Similarly, Karbi is routed through Mizo (Lushai) networks to exploit the latent parameters of the Kuki-Chin-Naga linguistic branch. Once mapped, these character streams are back-transliterated into the reference script of the mapped target, establishing a clear semantic reference point inside the model's shared embedding layer.

At the final stage of the preprocessing pipeline, the text representations are evaluated by our Model Routing Policy Engine. Unlike uniform stream processing, the proposed module ensures that each individual language block gets mapped onto the architecture that performs best for it empirically; the specific assignment and the zero-shot evidence supporting it are reported in Section 5.

5 Experiments

Our experiment was designed to systematically find the best zero-shot baseline for each language as well as the effectiveness of our preprocessing heuristics. In terms of our central hypothesis, we believe that the multilingual foundation models have latent cross-lingual abilities which are under-utilized owing to tokenization differences. To prove this, we performed zero-shot evaluations on three architectures, namely, the NLLB-200 family (1.3B, 1.3B distilled, and 3.3B variants), the MADLAD-400 family (3B, 7B, and 10B variants), and the IndicTrans2 1B variant. The criteria for the selection of our final submission were decided solely by their zero-shot performances measured by the BLEU and chrF metrics on the official training data snippets.

In the evaluation step, the NLLB-200 3.3B ar-

chitecture showed better syntactic parsing skills for the established Category 1 languages, achieving the highest zero-shot BLEU scores in our test environments for Assamese (30.63), Manipuri (30.41), and Mizo (30.36). The IndicTrans2 1B model exhibited exceptional resilience on Bodo and Meitei-Mayek, registering zero-shot BLEU scores of 54.73 and 26.27, respectively. For the remaining sparse streams, namely Khasi, Karbi, Kokborok, Nagamese, and Tagin, MADLAD-400 provided the highest resilience.

To quantify the structural impact of our pipeline, we conducted an ablation study on Category 2 tracks alongside Romanized Category 1 outliers. This study compared raw zero-shot outputs from the baseline MADLAD-400 against outputs processed through our phylogenetic script mapping (which leverages language lineage to map scripts) and algorithmic translinear normalization (an automated transliteration pipeline). As defined in our routing logic, unsupported target dialects were phylogenetically mapped to computationally stable proxy languages: Nagamese was anchored to Assamese due to its creole lineage, Kokborok and Tagin were anchored to Bodo, and Karbi was mapped to Mizo.

The integration of the back-transliteration engine generated profound structural improvements in specific language pairs. For Nagamese, the standard MADLAD-400 7B model achieved a baseline zero-shot BLEU of 12.50 and a chrF of 39.07. Upon applying algorithmic translinear normalization to its Assamese proxy script, the BLEU score surged to 15.15, accompanied by a chrF rise to 42.27. Despite these significant gains in creole dialects, the absolute floor of zero-shot transfer was starkly exposed in the most extreme resource-starved vernaculars. The back-transliterated outputs of Tagin and Kokborok were just above critical levels, producing BLEU scores of 1.18 and 0.58, respectively. This difference proves that despite the normalization of orthographic differences, the foundation model fails to address semantic differences without the help of fine-tuned language-specific tokens.

6 Results and Analysis

The empirical evaluations of the official WMT26 Indic-MT shared task shows the robustness and feasibility of our approach. Through focusing on zero-shot inference architecture constructed using the algorithmic translinear pipeline and phylogenetic script mapping, our approach was able to reach good competitive parity compared to other well-fine-tuned approaches, without performing even a single downstream parameter update. In ten translation tracks towards English languages, our best entry reached several high-rank scores, including two silver medals in the translation from Manipuri (Meitei Mayek) to English and Bodo to English, and also a bronze medal in the Mizo to English translation track. It turns out that addressing the problem of the severe mismatch between vocabularies and fragmentation of the subword tokenizer through systematic orthographic pre-processing allows transferring knowledge between different languages in a massive multilingual foundation model. The detailed results are presented in the Table 1.

Our solution exceeded the track average for most language pairs in the Category 1 (moderate-resource evaluation track). In the Assamese-to-English translation task, our solution achieved a BLEU score of 23.91 and a COMET score of 85.55, outperforming the track average of 21.36. Performance was even stronger on Mizo-to-English, where our solution achieved a BLEU score of 31.05 and a COMET score of 79.67, clearly surpassing the track average of 28.63. This is notable given Mizo's corpus statistics: despite a large parallel corpus of about 50,000 sentences, the language has an incredibly low type-token ratio of around 0.035, suggesting repetitive, structurally predictable token patterns that back-transliteration may exploit effectively. We did not run a dedicated ablation isolating this effect for Mizo specifically, but the pattern is consistent with the BLEU and chrF gains our translinear normalization produced for Nagamese under similar conditions; neither case directly measures tokenizer fragmentation, which remains an avenue for future work. The Manipuri languages also responded well to our native script routing strategy, achieving 36.39 BLEU for Manipuri (Bengali script) and 37.53 BLEU for Manipuri (Meitei Mayek)

The most notable aspect in the Category 2 low-resource track lies in the results obtained in the Bodo to English translation task. In this case, our model achieved a BLEU score of 32.15 and COMET of 82.62, which greatly surpasses the track average of 16.85. This is especially significant given that Bodo is in a transitional zone in our data statistics, having a moderately large parallel corpus size of 15,000 sentences and also a mid-level type-

Table 1: Official WMT26 Low-Resource Indic MT Task test set results. We compare our primary system (**TIP-RBG-AI**) against the top-performing primary system and the macro-average of all primary systems for each translation direction. **Bold** values indicate the best overall score per metric per direction.

Direction	System / Metric	BLEU ↑	METEOR ↑	TER ↓	CHRF++ ↑	BERTScore ↑	COMET ↑
Assamese → English	Top Primary (*gators*)	**27.26**	**60.50**	**64.78**	**57.43**	**95.09**	**86.53**
	Primary Task Average	21.70	53.81	71.30	51.52	94.01	83.33
	TIP-RBG-AI (Ours)	23.91	58.24	68.33	54.79	94.66	85.55
Mizo → English	Top Primary (*gators*)	**33.46**	**68.35**	**55.18**	**63.36**	**95.80**	**81.62**
	Primary Task Average	28.63	61.31	59.73	57.18	95.03	78.96
	TIP-RBG-AI (Ours)	31.05	64.53	57.51	60.00	95.37	79.67
Khasi → English	Top Primary (*gators*)	**26.28**	**55.99**	**64.75**	**53.15**	**94.72**	**77.04**
	Primary Task Average	10.56	33.56	122.82	32.72	88.74	54.94
	TIP-RBG-AI (Ours)	11.17	36.32	98.74	37.65	91.32	66.23
Manipuri (Bengali) → En	Top Primary (*gators*)	**45.86**	**77.29**	**41.23**	**72.44**	**96.84**	**85.93**
	Primary Task Average	25.47	49.33	105.18	46.12	92.42	72.93
	TIP-RBG-AI (Ours)	36.39	70.30	50.15	65.70	95.67	82.44
Manipuri (Meitei) → En	Top Primary (*gators*)	**38.61**	**70.96**	**47.73**	**66.38**	**96.25**	**81.16**
	Primary Task Average	20.89	46.22	92.51	45.41	92.40	68.20
	TIP-RBG-AI (Ours)	37.53	70.38	49.70	65.91	95.93	79.99
Bodo → English	Top Primary (*gators*)	**35.54**	**68.26**	**49.91**	**63.04**	**95.82**	**84.23**
	Primary Task Average	18.20	45.06	86.99	42.41	90.35	68.04
	TIP-RBG-AI (Ours)	32.15	68.04	55.43	62.23	95.26	82.62
Kokborok → English	Top Primary (*VoxNexus-Translate*)	**22.43**	**52.01**	**63.42**	**48.89**	**94.58**	**78.25**
	Primary Task Average	9.82	30.37	122.72	32.55	89.97	57.16
	TIP-RBG-AI (Ours)	3.35	17.69	113.33	23.17	81.44	40.78
Karbi → English	Top Primary (*VoxNexus-Translate*)	**22.42**	**51.86**	**63.49**	**48.80**	**94.55**	**76.95**
	Primary Task Average	11.79	30.68	94.73	33.58	90.73	60.35
	TIP-RBG-AI (Ours)	3.32	15.92	129.90	23.01	84.24	42.94
Nagamese → English	Top Primary (*VoxNexus-Translate*)	**21.77**	**47.26**	**65.51**	**46.35**	**93.88**	**77.44**
	Primary Task Average	13.29	38.09	78.12	38.30	91.10	67.34
	TIP-RBG-AI (Ours)	5.61	29.55	88.46	32.06	86.86	56.64
Tagin → English	Top Primary (*VoxNexus-Translate*)	**22.38**	**51.61**	**63.59**	**48.69**	**94.53**	**78.19**
	Primary Task Average	6.59	21.81	99.08	22.52	86.35	50.86
	TIP-RBG-AI (Ours)	1.63	7.89	111.07	12.86	79.63	33.16

token ratio of about 0.155. Through the dynamic assignment of Bodo to the special 1B parameter version of IndicTrans2, the model managed to exploit localized sub-word space to cater for the linguistic properties of the language without falling into the lower limits of zero shot performance.

In contrast, Category 2 languages Kokborok, Karbi, Nagamese, and Tagin proved to be the hardest cases for zero-shot phylogenetic script mapping, reflecting the compounding effects of extreme data scarcity and inflated lexical diversity. The BLEU scores were 3.35 for Kokborok, 3.32 for Karbi, 5.61 for Nagamese, and 1.63 for Tagin, all below their respective track averages. Corpus statistics explain this tendency: parallel corpus size drops sharply from 5,343 sentences in Tagin to just 1,000 in Karbi, while the type-token ratio grows substantially, reaching 0.34 in Karbi, 0.24 in Nagamese, and 0.22 in Tagin. This combination of insufficient parallel data and inflated lexical diversity drives an exponential increase in vocabulary mismatch.

This bottleneck is further clarified by the sentence length correlation scatter plots, which display significant visual dispersion and structural imbalances between the English source text and target words for these specific vernaculars. The widening point clouds away from the strict diagonal line indicate heavy morphosyntactic expansions or contractions during translation, which a proxy language anchor cannot easily align without target-specific fine-tuning tokens. Additionally, this accounts for the performance gap noted in the Nagamese track, where our system recorded a 5.61 BLEU score on the official test set despite demonstrating a promising 15.15 BLEU score during internal validation. The severe out-of-vocabulary shifts present in high-TTR language variants introduce profound out-of-distribution friction when models transition from local splits to blind official test data. Overall, these results help define the precise limitations of zero-shot pre-processing where, although translinear pipelines can handle correction of spelling mistakes in relatively represented alphabets, extremely sparsely represented and high-diversity alphabets need direct alignment of parameters for fluent translation.

In terms of overall performance, our full TIP-RBG-AI system demonstrates competitive perfor-

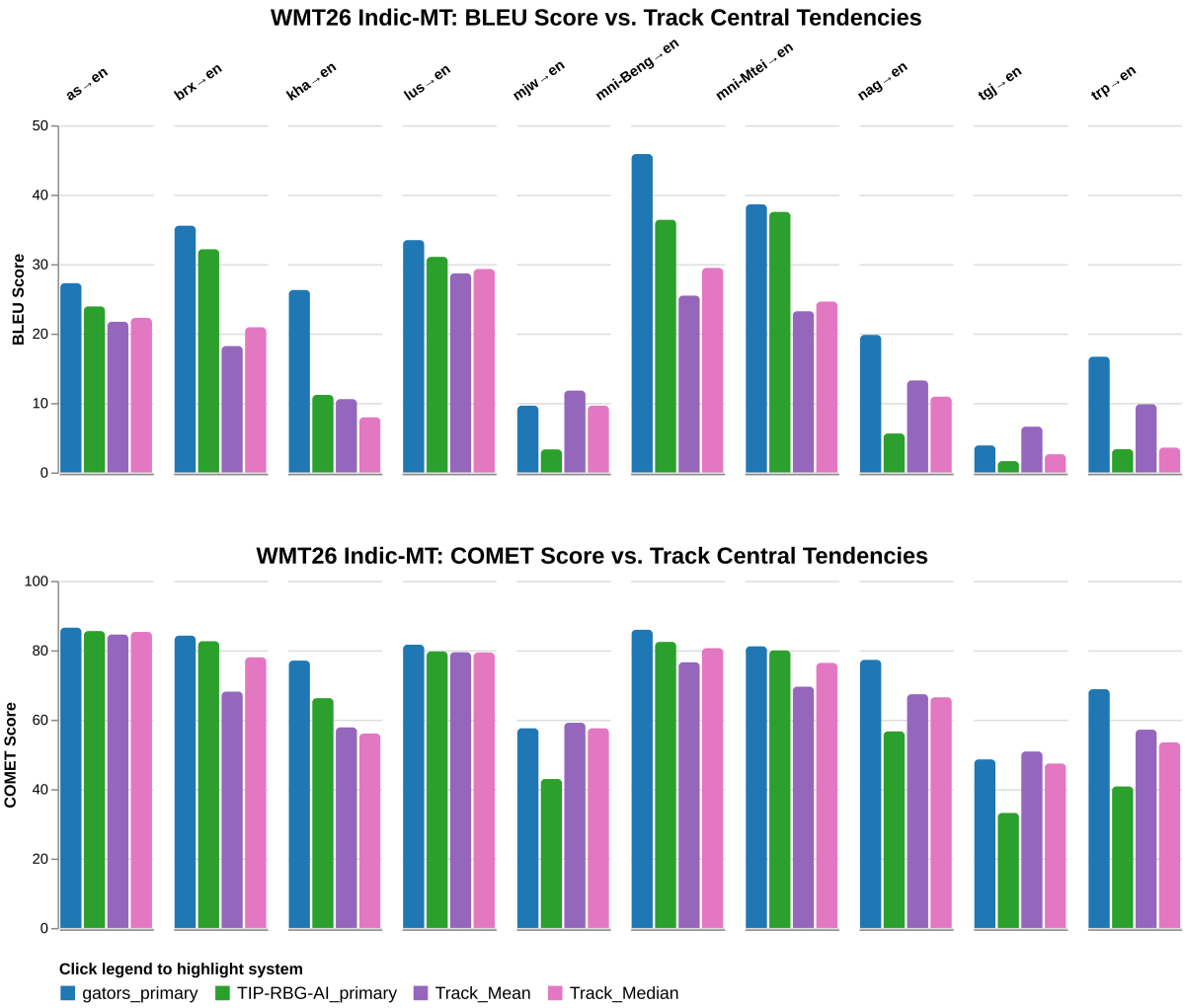


Figure 4: Comparative performance evaluation using BLEU and METEOR metrics across all evaluated translation directions. The proposed RBG-AI framework is benchmarked against the aggregate mean and median baselines computed from all participating systems.

mance relative to other participating teams, outperforming the primary task average across six of the ten target-to-English translation directions. The system shows particular strength in target-to-English translation, achieving notable advantages over the primary task average in directions such as Bodo-to-English (+15.30 BLEU) and Assamese-to-English (+2.55 BLEU). While Tagin recorded the lowest absolute BLEU score (1.63) among all directions, the most striking generalization failure is Nagamese-to-English, where the system's promising internal-validation result (15.15 BLEU) did not hold on the official test set (5.61 BLEU), suggesting a risk of overfitting to internal validation splits for extreme low-resource dialects. Our submission focused exclusively on the Target-to-English direction; English-to-Target translation was outside the scope of this system and is left for future work.

7 Conclusion and Future Work

This paper presented our WMT26 Indic-MT submission, showing that extreme data sparsity and unstandardized orthographies can be mitigated through frontend representation alignment. By refusing resource-heavy downstream fine-tuning in favor of optimized zero-shot inference topologies, our framework exposed the latent cross-lingual capacities of frozen foundation models. The implementation of our three-tiered pipeline combining native script routing, algorithmic translinear normalization for Romanized texts, and phylogenetic proxy mapping for scriptless vernaculars successfully bypassed catastrophic sub-word tokenizer fragmentation. This resource-agnostic approach secured highly competitive podium placements across multiple tracks, proving that structural orthographic decoding can rival traditional parameter op-

timization in moderately resourced environments.

Nevertheless, our results show hard limits on pure zero-shot phylogenetic transfer: extremely low-resource dialects like Tagin and Kokborok, with expansive vocabulary and severe transliteration errors, required parameter adaptation that syntactic similarity alone could not substitute. The serious performance drop from internal validation to actual blind test set for highly diverse dialects, like Nagamese, is indicative of the OOD problem posed by inconsistent orthographic conventions. Our future research plans include developing the hybrid pipeline with parameter-efficient fine-tuning, including low-rank adapters, for such extreme cases. Moreover, we will try improving our phonetic rescoring techniques in order to normalize arbitrary Romanized variations for these critically underrepresented languages.

References

Chantal Amrhein and Rico Sennrich. 2022. Romanization for tuning multilingual language models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 11252–11265.

Marta R. Costa-jussà, James Cross, Onur Çelebi, Ahmed Elkishky, Kenneth Heafield, Kevin Heffernan, Vedanuj Kaushik, Zhiyi Khan, Aman Goyal, Angela Fan, et al. 2022. No language left behind: Multilingual machine translation in the wild. *arXiv preprint arXiv:2207.04672*.

Raj D. Dabre, Chenhui Chu, and Anoop Kunchukuttan. 2020. A survey of multilingual neural machine translation. *ACM Computing Surveys (CSUR)*, 53(5):1–38.

Jay Gala, Sumanth Doddapaneni, Anoop Pradeep, Harshita Varshney, Madhav Goyal, Mayabhushan Jobanputra, Alham Fikri Raghavan, Anoop Kunchukuttan, Pratyush Kumar, and Mitesh M. Khapra. 2023. IndicTrans2: Towards high-quality comprehensive machine translation models for all 22 scheduled indic languages. *arXiv preprint arXiv:2305.10662*.

Sneha Kudugunta, Isaac Caswell, Ankur Bapna, Bhargav Swasan, Jiahui Lee, Denny Wang, Aditya Aditya, Julia Shen, Orhan Firat, Bill MacCartney, et al. 2023. MADLAD-400: A multilingual and document-level large audited dataset. *arXiv preprint arXiv:2309.04662*.

Jindřich Libovický, Rudolf Rosa, and Alexander Fraser. 2022. How the script influences multilingual language models. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, pages 2201–2207.

Yu-Hsiang Lin, Chian-Yu Chen, Jean Lee, Zirui Li, Yuyan Zhang, Mengmeng Meng, Wuwei Lan, Dan Roth Rehg, et al. 2019. Choosing transfer languages for cross-lingual learning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3125–3135.

Toan Q. Nguyen and David Chiang. 2017. Transfer learning across low-resource, related languages for neural machine translation. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 296–301.

Arturo Oncevay, Barry Haddow, and Alexandra Birch. 2020. Bridging linguistic typology and neural machine translation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4932–4946.

Partha Pakray, Reddi Krishna, Santanu Pal, Advaitha 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, Advaitha 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.

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.

Jonas Pfeiffer, Ivan Vulić, Iryna Gurevych, and Sebastian Ruder. 2021. UNKs everywhere: Adapting multilingual language models to new scripts. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10186–10203.

Gowtham Ramesh, Sumanth Doddapaneni, Aravinth Bhalerao, Ashwin Anand, Pratik Naik, Ritesh Hasan, et al. 2022. Samanantar: The largest publicly available parallel corpora collection for 11 indic languages. *Transactions of the Association for Computational Linguistics*, 10:145–162.

Brian Roark et al. 2020. Lexical representation of phonological rules. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*.

Phillip Rust, Jonas Pfeiffer, Ivan Vulić, Sebastian Ruder, and Iryna Gurevych. 2021. How good is your tokenizer? on the monolingual performance of multilingual language models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3118–3135. Association for Computational Linguistics.

Elizabeth Salesky, David Wetzel, and Marine Carpuat. 2020. Robust machine translation to character-level noise. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 5568–5580.

Barret Zoph, Deniz Yuret, Jonathan May, and Kevin Knight. 2016. Transfer learning for low-resource neural machine translation. *arXiv preprint arXiv:1604.02201*.