

SCE-KIIT: Fine-Tuning NLLB-200 for English-Nagamese Translation

Agniva Maiti[†] Aarsh Muley[†] Sovan Kumar Sahoo[†]

[†]School of Computer Science and Engineering

Kalinga Institute of Industrial Technology (KIIT), Bhubaneswar, India

2205964@kiit.ac.in 22051304@kiit.ac.in sovan.sahoofcs@kiit.ac.in

Abstract

Nagamese is a critically under-resourced Assamese-based creole language spoken in Nagaland, India, serving as the primary lingua franca for over two million speakers across 16 Naga tribal communities. Written in the Latin script, it is entirely absent from all modern massively multilingual machine translation (MT) systems. We describe the SCE-KIIT submission to the WMT 2026 Shared Task on Low-Resource Indic Language Translation (Category 2: English $\leftrightarrow$ Nagamese), for which we submitted a primary EN $\rightarrow$ NAG run. Our system builds on Facebook’s NLLB-200 distilled with 600M parameters using a two-phase surrogate-token strategy: we fine-tune with Assamese (`asm_Beng`) as the target-side language prefix, then force `asm_Latn` at inference time. This token is absent from NLLB’s vocabulary and resolves to `<unk>`; we hypothesise that seeding generation with an out-of-vocabulary token leaves the decoder’s script preference unconstrained, allowing the Romanised output learned during fine-tuning to surface. On our 228-sentence held-out test split, the system achieves SacreBLEU 40.02, chrF 53.61, and COMET 0.7212 (`wmt22-comet-da`), against a zero-shot baseline of 0.40 BLEU and 0.4487 COMET: an absolute gain of 39.62 BLEU. On the official shared-task test set, which is out-of-domain news text, the same system scores 12.51 BLEU and 43.81 chrF++; we analyse this domain gap in detail. The model and corpus are publicly released at <https://huggingface.co/agnivamaiti/nllb-200-en-nagamese>.

1 Introduction

The past decade has witnessed a sea change in machine translation, driven by the emergence of large-scale multilingual neural architectures. Models such as NLLB-200 (Costa-jussà et al., 2022), trained on web-crawled corpora spanning more than 200 languages, represent a genuine milestone

in the democratisation of language technology. Yet the very data-centric philosophy that powers these models carries within it a systemic blind spot: languages that exist primarily in speech, in handwritten registers, or in communities too small to generate a measurable digital footprint remain invisible to them. For such languages, the promise of multilingual NLP is largely illusory (Ranathunga et al., 2023).

Nagamese sits squarely in this excluded category. It is an Assamese-based creole native to the northeastern Indian state of Nagaland, where it functions as the essential medium of inter-tribal communication, commerce, and administration for a population exceeding two million across 16 major Naga tribes, each speaking a distinct Tibeto-Burman mother tongue. Historically, Nagamese emerged in the 19th century as a contact pidgin between Assamese merchants and Naga communities, and over successive generations creolised into a fully-formed language with an Indo-Aryan lexicon layered atop Tibeto-Burman syntactic and phonological structures. Despite its vital sociolinguistic role, Nagamese has no standardised orthography and no representation in any multilingual MT system. Nor, prior to this work, did it have a curated parallel corpus suitable for MT: written Nagamese exists online in scattered and unaligned form—scripture translations, learner phrasebooks, regional news, and community pages—from which we assemble and align the corpus described in Section 3.

The practical challenge is compounded by a critical script divergence. Standard Assamese, the closest relative of Nagamese, is written in the Assamese-Bengali script and holds a dedicated language code (`asm_Beng`) in NLLB-200. Written Nagamese, however, has adopted the Latin (Roman) script, driven by the widespread use of English in the Nagaland education system. This single fact undermines the most natural baseline strat-

egy: using `asm_Beng` as a proxy for Nagamese causes NLLB’s script-forcing mechanism to generate Bengali-character output that is phonetically and visually unintelligible to Nagamese readers, yielding a SacreBLEU score of approximately zero against Romanised references. Conversely, selecting a Latin-script code such as `eng_Latn` or `hin_Latn` abandons the Assamese morphological and lexical priors that would make transfer possible, resulting in fluent but semantically disconnected output.

In this paper, we present a targeted solution to this dual challenge and describe the SCE-KIIT system submitted to the WMT 2026 Low-Resource Indic Language Translation Shared Task (Category 2). Our approach is built around a two-phase surrogate-token strategy: during fine-tuning on our curated parallel corpus, we expose the decoder to `asm_Beng` as its target prefix, grounding it in NLLB’s Assamese representation space; at inference time, we force the `asm_Latn` token, which is absent from NLLB’s vocabulary and is therefore mapped by the tokeniser to `<unk>` (ID 3). Our working hypothesis is that this leaves the decoder’s script preference unconstrained rather than actively overriding it; we did not run the ablation that would separate the two explanations, and return to this in the Limitations section. Empirically, the result is a decoder that retains the morphological richness of Assamese-derived cross-lingual representations while generating Romanised Nagamese exactly as seen during training.

A model that achieves 0.40 BLEU in the zero-shot setting rises to 40.02 BLEU after fine-tuning on fewer than 2,000 sentence pairs, an absolute gain of 39.62 BLEU. Our three primary contributions are as follows. First, we curate, clean, and publicly release a parallel corpus of 2,280 English–Nagamese sentence pairs, which constitutes the largest openly available aligned resource for this language. Second, we formalise the two-phase surrogate-token strategy as a replicable methodology for adapting multilingual MT systems to Romanised creole languages with no native model support. Third, we provide a comprehensive automatic evaluation using SacreBLEU (Post, 2018), chrF (Popović, 2015), and COMET (Rei et al., 2020), alongside a qualitative analysis demonstrating the system’s ability to handle domain-specific named entities, idiomatic Nagamese phrasing, and morphological agreement.

2 Related Work

2.1 Low-Resource Machine Translation

The challenge of translating under conditions of severe data scarcity has a long history in the field. Statistical MT systems (Koehn et al., 2007) were particularly sensitive to sparsity due to their dependence on phrase alignment tables, and neural MT models (Vaswani et al., 2017), despite their expressive power, are notoriously data-hungry. The community has responded with a toolkit of complementary approaches: subword tokenisation via Byte-Pair Encoding (Sennrich et al., 2016) to reduce out-of-vocabulary rates, data augmentation through back-translation (Edunov et al., 2018), cross-lingual transfer from high-resource relatives (Zoph et al., 2016), and unsupervised cross-lingual pre-training (Lample and Conneau, 2019). A comprehensive survey by Ranathunga et al. (2023) establishes that combining large-scale multilingual pre-training with targeted fine-tuning is the current state of the art, which is precisely the paradigm we adopt here.

2.2 Multilingual Pre-trained Models

The massively multilingual pre-training paradigm began with encoder-only models such as mBERT (Devlin et al., 2019) and XLM-R (Conneau et al., 2020) and matured into full sequence-to-sequence architectures including mT5 (Xue et al., 2021), mBART (Tang et al., 2020; Lewis et al., 2020), and NLLB-200 (Costa-jussà et al., 2022). These models induce a shared multilingual representation space that facilitates zero-shot and few-shot cross-lingual transfer (Johnson et al., 2017; Aharoni et al., 2019). However, scaling to hundreds of languages simultaneously introduces capacity competition and negative interference, the “curse of multilinguality” (Fan et al., 2021), which motivates the targeted fine-tuning approach we pursue rather than zero-shot application.

2.3 Indic Language Translation

Systematic efforts to build MT resources for Indian languages have produced large-scale parallel corpora such as Samanantar (Ramesh et al., 2022) and evaluation benchmarks such as Indic-NLPSuite (Kakwani et al., 2020), though coverage remains concentrated on the 22 constitutionally scheduled languages. The WMT Low-Resource Indic Language Translation Shared Task has progressively expanded coverage to regional minority

languages, with prior editions addressing English–Nyishi (Kakum et al., 2023) and other underserved pairs (Pal et al., 2023; Pakray et al., 2024, 2025a,b). The WMT 2026 edition introduces Nagamese as a new target language, providing the impetus for the present work.

2.4 Surrogate Language and Script Adaptation

The principle of bootstrapping an unsupported language through a typologically or genetically related surrogate is well-established in low-resource NLP (Neubig and Hu, 2018). For Nagamese, Assamese is the natural genetic surrogate given shared lexical roots and morphological patterns. The key novelty of our approach is the decoupling of the training-time surrogate (which anchors the representation space) from the inference-time surrogate (which we argue leaves script selection unconstrained). This two-phase design extends prior surrogate language work to the specific challenge of script mismatch, a setting not directly addressed by existing methods. Future work will examine whether parameter-efficient fine-tuning via LoRA (Hu et al., 2021) can scale this strategy to the full NLLB-3.3B checkpoint.

3 Dataset and Preprocessing

3.1 Data Sources

We use the `nagamesmt` parallel corpus, comprising English–Nagamese sentence pairs drawn from biblical translations, conversational phrasebooks, regional news sources, and community web pages (<https://huggingface.co/datasets/agnivamaiti/nagamese-english-mt>). After preprocessing, the corpus contains 2,280 high-quality aligned pairs. Using a fixed random seed (`random_state=42`) for full reproducibility, we partition the data according to the split statistics reported in Table 1. The public release of this corpus has since grown to 4,178 pairs (3,342/418/418) through continued collection. All results in this paper are computed on the earlier 2,280-pair snapshot described above; we report the snapshot used rather than the current release so that the figures in Table 3 remain reproducible.

3.2 Preprocessing and Filtering

Creole text crawled from heterogeneous web sources frequently exhibits noisy formatting and orthographic variation. We apply a four-stage clean-

Split	Pairs	Share
Train	1,824	80%
Validation	228	10%
Test	228	10%
Total	2,280	100%

Table 1: Corpus split statistics (`random_state=42`).

ing pipeline. All sentences are first converted to Unicode NFC form to resolve encoding anomalies. Curly quotation marks and typographic apostrophes are then replaced with standard ASCII equivalents to reduce tokeniser vocabulary fragmentation. Next, sentence pairs containing fewer than three or more than 300 tokens on either side are discarded to eliminate trivial fragments and GPU memory overflows. Finally, pairs whose word-count ratio exceeds 4.0 or falls below 0.25 are removed as likely misalignments. All 2,280 retained pairs satisfy every constraint, confirming the corpus’s alignment quality.

4 Methodology

4.1 Base Model Architecture

Our system is built on Facebook’s NLLB-200 distilled with 600M parameters (Costa-jussà et al., 2022), a knowledge-distilled variant of the 3.3B-parameter NLLB model that retains high multilingual translation quality while fitting within the 16 GB VRAM of a single NVIDIA Tesla T4 GPU. The architecture comprises 12 encoder layers, 12 decoder layers, 16 attention heads per layer, and a hidden dimension of 1,024.

4.2 The Surrogate-Token Strategy

The absence of a dedicated NLLB language code for Nagamese forces a design choice about which existing code to use as a proxy. Selecting `asm_Beng` at inference time activates NLLB’s internal script-forcing mechanism, which constrains the decoder to emit Bengali-script characters, output that is phonetically and orthographically incompatible with Romanised Nagamese references. Selecting a Latin-script code such as `eng_Latn` abandons the Assamese morphological priors that would enable effective transfer. Our two-phase strategy sidesteps both traps.

During the training phase, the decoder’s forced beginning-of-sequence (BOS) token is set to `asm_Beng`. This anchors the model’s representation space to NLLB’s pre-trained Assamese

parameters—which share extensive lexical and morphological overlap with Nagamese—while the cross-entropy loss on the Latin-script target corpus simultaneously adapts the decoder’s projection head to map these representations onto Romanised character sequences.

During the inference phase, the BOS token is forced to `asm_Latn`. Because `asm_Latn` is absent from NLLB’s native vocabulary, the tokeniser maps it to the unknown token (`<unk>`, ID 3)—confirmed by `tokenizer.convert_ids_to_tokens(3)`. We interpret this single-token substitution as a surrogate bypass, in which the script-forcing gate that would otherwise lock the decoder into Bengali-script output is never triggered, leaving the fine-tuned model free to decode the Romanised Nagamese it learned during training. We emphasise that the tokeniser mapping is verified whereas this interpretation is not: the observable fact is that decoding under `asm_Latn` yields fluent Romanised output, and the Limitations section states the ablation needed to establish the mechanism. The two pipelines are illustrated in Figure 3.

4.3 Fine-Tuning Configuration

We fine-tuned the model using the Hugging Face `Seq2SeqTrainer` for 15 epochs with full-parameter optimisation (no LoRA or adapter layers). Both translation directions (EN→NAG and NAG→EN) are trained jointly by prefixing each batch with the appropriate source and target language tokens. We stress that joint training does not imply joint evaluation: every result in this paper is EN→NAG, and EN→NAG is the only direction we submitted to the shared task. The reverse direction was trained but never systematically evaluated, and we make no claim about its quality (see the Limitations section). Table 2 summarises all hyperparameters and the hardware configuration.

5 Experimental Setup

5.1 Evaluation Metrics

We report four complementary metrics:

1. **SacreBLEU** (Post, 2018): Word-level BLEU score.
2. **chrF** (Popović, 2015): Character n -gram F-score, particularly robust to spelling variations in creole text.

Setting	Value
Base checkpoint	NLLB-200 distilled (600M)
Optimiser	AdamW ($\lambda=0.01$)
Learning rate	5×10^{-5} , linear warm-up (500 steps)
Micro-batch / accum.	2 / 4 (effective batch: 8)
Training epochs	15
Precision	FP16
Beam size	5
<code>no_repeat_ngram</code>	3
<code>length_penalty</code>	1.0
<code>max_length</code>	128 tokens
Early stopping	best validation BLEU
GPU	Tesla T4 (15.6 GB VRAM)
CUDA / PyTorch	13.0 / 2.11.0

Table 2: Fine-tuning hyperparameters and hardware.

3. **TER** (Snover et al., 2006): Translation Edit Rate (lower is better), which measures the minimum number of edit operations (insertions, deletions, substitutions, and shifts) required to transform the system output into the reference.
4. **COMET** (Rei et al., 2020): A neural translation evaluation metric. We use the Unbabel’s WMT22 COMET-DA model, which correlates strongly with human judgements of translation quality.

5.2 Baseline

We compare against zero-shot translation with the unmodified Facebook’s NLLB-200 distilled with 600M parameters checkpoint. To ensure a fair comparison of the models’ actual linguistic competence, the zero-shot baseline uses the identical `asm_Latn` inference bypass, so both conditions generate Latin-script output rather than Bengali-script text.

6 Results and Analysis

6.1 Quantitative Evaluation

Table 3 reports the evaluation results on the 228-sentence held-out test split. Our fine-tuned system substantially outperforms the zero-shot baseline across all four metrics.

The fine-tuned model improves SacreBLEU by 39.62 points, chrF by 40.60 points, and COMET by 0.2725 points. We report these as absolute differences rather than as a ratio: the zero-shot baseline is close enough to zero that a relative figure would be dominated by noise in the denominator. A COMET score of 0.7212 is particularly telling: this neural metric evaluates semantic adequacy and

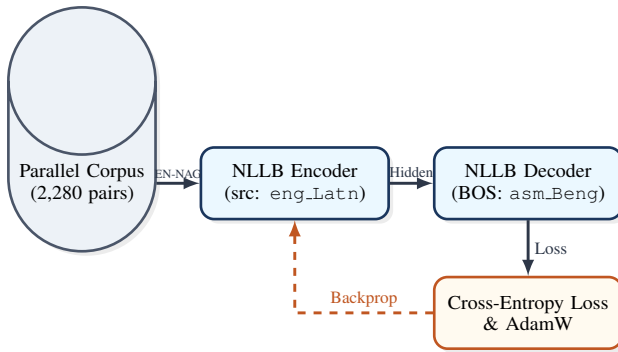


Figure 1: (a) Training: `asm_Beng` as decoder BOS anchors Assamese representation space.

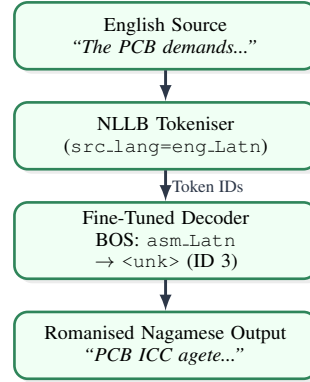


Figure 2: (b) Inference: `asm_Latn` $\rightarrow$ `<unk>`, which we hypothesise leaves the script-forcing gate untriggered.

Figure 3: System architecture. Training (a) uses `asm_Beng` to anchor Assamese priors; inference (b) forces `asm_Latn`, which the tokeniser maps to `<unk>` (ID 3). The model generates fluent Romanised Nagamese under this setting; the proposed mechanism, that the script-forcing gate is left untriggered, is a hypothesis we do not isolate experimentally.

System	BLEU $\uparrow$	chrF $\uparrow$	TER $\downarrow$	COMET $\uparrow$
NLLB-200-600M (Zero-Shot)	0.40	13.01	–	0.4487
NLLB-200-600M (Fine-Tuned)	40.02	53.61	50.49	0.7212
Δ	+39.62	+40.60	–	+0.2725

Table 3: System comparison on the internal held-out test split ($n = 228$, EN $\rightarrow$ NAG). Δ shows absolute improvement of our fine-tuned system. TER is not reported for the zero-shot baseline owing to near-zero overlap with references.

fluency jointly, indicating that our translations are not only superficially correct but are also meaningfully aligned with the reference in content and style. The size of these gains, achieved with only 1,824 training sentences, highlights both the strength of the surrogate-token transfer and the effectiveness of NLLB’s pre-trained Assamese representations as a starting point for Nagamese adaptation.

While our translation quality is high, the Translation Edit Rate (TER) for our fine-tuned model is relatively high at 50.49. We attribute this high error rate to several key factors. First, Nagamese is primarily an oral language with no official standardized orthography or spelling system. As a result, spelling is highly inconsistent across speakers, writers, and the web-crawled training/test corpus. This leads to discrepancies between our correct, fluent translations and the specific references (e.g., lexical choices like *ami* vs. *Moi* for “I”, or *ya* vs. *Hoy* for “yes”), which are heavily penalized by word-level edit distance. Second, our preprocessing and filtering pipeline, while necessary to reduce tokenizer

vocabulary fragmentation, may occasionally introduce tokenization artifacts or strip minor phonetic nuances. These factors combine to increase the edit-based error rate, despite the model producing semantically accurate and fluent outputs.

6.2 Official Shared-Task Results

Table 4 reports the organisers’ evaluation of all primary EN $\rightarrow$ NAG submissions on the official 1,000-sentence test set (Pakray et al., 2026). The organisers report per-metric scores and do not declare a ranking or a winning system; systems are listed here in the descending BLEU order in which they were released. Our primary run placed fourth of five primary submissions.

System	BLEU $\uparrow$	chrF++ $\uparrow$	COMET $\uparrow$
Decepticons-IIITH	20.34	57.66	69.55
gators	17.44	54.93	66.34
ANVITA	16.37	51.81	65.16
SCE-KIIT (ours)	12.51	43.81	59.04
KrenTransV0-1	0.67	16.40	44.16

Table 4: Official EN $\rightarrow$ NAG results on the shared-task test set ($n = 1,000$), primary systems only. COMET is reported on a 0–100 scale. Our run additionally scores METEOR 42.54, TER 91.82, and BERTScore 77.45.

The gap between these figures and those in Table 3 is large and deserves a direct explanation: 12.51 BLEU on the official test set against 40.02 on our internal split. The two numbers measure different things. Our internal split is a random partition of the same curated corpus used for training, and

is therefore in-domain: predominantly scriptural and conversational material, with the orthographic conventions of the sources we aligned. The official test set is contemporary news text, a domain absent from our training data, and one dense in named entities, organisations, and recently coined terminology. The 27.5-point drop is a measure of domain transfer, not of measurement error, and we regard the official figure as the better estimate of this system’s practical utility. Section 6.4 examines the submitted outputs directly and identifies where the loss occurs.

We note two further points of comparability. The organisers report chrF++ whereas Table 3 reports chrF, and they report COMET on a 0–100 scale against our 0–1 scale; our internal COMET of 0.7212 corresponds to 72.12 in the units of Table 4. All systems in Table 4 are evaluated on identical data by the organisers, so the relative ordering is directly comparable even where our internal figures are not. More generally, cross-corpus comparison in Nagamese is complicated by the absence of a standard written norm: independently assembled corpora differ in orthography, register, and degree of Assamese versus English lexical borrowing, and these differences affect every system evaluated on a given test set equally.

6.3 Qualitative Analysis

Table 5 presents representative translation outputs from the fine-tuned system. The examples demonstrate the model’s ability to handle everyday conversational text, complex sentence structures, and culturally specific terms, while exhibiting natural orthographic variations.

In Example 1, the model translates the English source with high fluency, using the natural Nagamese conjunction *kilemane* (because) and the construction *kisa bisi dukh thakise* (the story was very sad), which matches the meaning perfectly, albeit with slight syntactic variation from the reference. In Example 2 and Example 3, the predictions differ slightly from the reference by substituting common synonyms: *ami* for the first-person singular pronoun instead of the reference’s *Moi*, and the informal agreement marker *ya* (yes) instead of the reference’s formal *Hoy*. These natural variations represent correct translations but explain the high edit rate captured by the TER metric. Lastly, Example 4 displays the model’s ability to generate exact translations for culturally specific items, correctly mapping “bamboo shoot” to *baastenga* and “pork”

to *gahori mangsho*, reflecting a deep grounding in regional vocabulary.

6.4 Analysis of the Submitted Test Run

The examples above are drawn from our internal split, where references are available. The official shared-task test set is released source-side only, so we cannot compute reference-based error rates on it. We can, however, characterise the 1,000 submitted EN→NAG outputs directly, which tells us where the 27.5-point BLEU drop reported in Section 6.2 comes from. Table 6 summarises the run.

Three observations follow. First, the system is well-behaved under domain shift: no output is empty, only four show intra-sentence trigram repetition, and only four remain near-identical to their source. Degenerate decoding, a common failure mode when a fine-tuned model meets out-of-domain input, is essentially absent.

Second, 36.0% of all output tokens are carried over from the English source. This is not primarily a translation failure. Of those, 23.5 percentage points are capitalised tokens—named entities, organisations, and acronyms such as *BCCI*, *ICC*, and *PCB*—which Nagamese conventionally leaves in English. The remaining 12.6 points are lowercase English loanwords (*match*, *schedule*, *title*), consistent with the pervasive code-mixing of contemporary written Nagamese. The output/source length ratio of 1.12 indicates the model is producing full translations rather than truncating or copying.

Third, the run exposes a concrete failure mode invisible in Table 5: corruption of rare proper nouns. The surname *Viswanathan* is rendered correctly in three of five occurrences but becomes *Worldwanathan* in one and *Wiswanathan* in another; in a further case an ordinal is truncated, leaving an output that begins “*th edition te...*”. The first of these is diagnostic: the model has segmented an unseen name into subwords and translated the fragment *Vis*→*World*, applying lexical transfer where verbatim copying was required. Named-entity handling is therefore the most promising target for improvement, and a copy mechanism or entity-aware constrained decoding would address it directly.

Finally, the run gives independent support to the orthographic explanation we advanced for the high TER in Section 6.1: the model itself uses competing forms for the same function within a single run. The copula is realised as *ase* 79% of the time but *thake* 14% and *ahe* 7% ($n = 99$); ‘do/make’ as *kore* 84%, *kuri* 14%, *kori* 2% ($n = 124$); and

Source (EN)	Sarah cried because the story of the movie was very sad.
Reference	Movie laga kisa bisi dukh thaka karon Sarah kandishe.
Prediction	Sarah kandi jaishe kilemane movie laga kisa bisi dukh thakise.
Source (EN)	I am feeling much better now.
Reference	Moi etiya bisi bhal feel kuri ase.
Prediction	ami etiya bisi bhal feel kuri ase.
Source (EN)	Yes, please sit.
Reference	Hoy, morom para bohobi.
Prediction	ya, morom para bohobi.
Source (EN)	We love eating bamboo shoot and pork.
Reference	Amikhan baastenga aru gahori mangsho khabole bhal pai.
Prediction	Ami khan baastenga aru gahori mangsho khabole bhal pai.

Table 5: Representative translation outputs from the fine-tuned system (EN→NAG).

Property	Value
Sentences submitted	1,000
Output tokens	12,785
Empty outputs	0
Outputs with a repeated trigram	4 (0.4%)
Near-identical to source	4 (0.4%)
Output/source length ratio	1.12
Output tokens copied from source	36.0%
capitalised (entities)	23.5%
lowercase (loanwords)	12.6%

Table 6: Reference-free properties of the submitted primary EN→NAG run on the official test set ($n = 1,000$).

the locative/ablative as *te* 78%, *pora* 16%, *logote* 6% ($n = 800$). When a language admits several spellings for the same morpheme, any single reference penalises the alternatives, and edit-distance metrics absorb that penalty in full.

6.5 Surrogate Token Mechanics

One might ask whether the `asm_Latn` bypass is principled or coincidental. We can state one part of this with certainty. Querying the NLLB-200 tokeniser directly, `convert_tokens_to_ids("asm_Latn")` returns ID 3, which `convert_ids_to_tokens(3)` confirms is `<unk>`; for comparison, `asm_Beng` returns 256015 and `eng_Latn` 256047 in a vocabulary of 256,204 entries. Passing `asm_Latn` as the forced BOS token therefore seeds generation with `<unk>` at every inference call. This is a verifiable property of the tokeniser, not an inference from our results.

The mechanism by which this produces Romanised output is, however, a hypothesis rather than a demonstrated fact. During fine-tuning, all training sequences began with `asm_Beng` as their BOS

token, yet the targets were Romanised Nagamese; we suggest that the decoder therefore associates its Assamese-anchored hidden states with Latin-character output, and that seeding with an unseen `<unk>` leaves the script preference unconstrained so that these learned patterns surface. An alternative explanation is that fine-tuning alone suffices and the choice of BOS token is largely immaterial. Distinguishing the two requires decoding the same fine-tuned checkpoint under `asm_Beng` and `asm_Latn` on the same test split, an ablation we did not run; we flag it in the Limitations section and do not claim the stronger reading here.

7 Conclusion

We presented SCE-KIIT’s submission to the WMT 2026 Low-Resource Indic Language Translation Shared Task (Pakray et al., 2026), Category 2 (English–Nagamese), a low-resource creole language with no prior MT support. By applying a two-phase surrogate-token strategy, using `asm_Beng` as the decoder BOS during fine-tuning and `asm_Latn` (`<unk>`, ID 3) at inference, we adapted the NLLB-200 distilled with 600M parameters model to generate fluent Romanised Nagamese. On our 228-sentence held-out test split, the EN→NAG system achieves SacreBLEU 40.02, chrF 53.61, and COMET 0.7212, compared to a zero-shot baseline of 0.40 BLEU and 0.4487 COMET. On the official shared-task test set it scores 12.51 BLEU, placing fourth of five primary submissions; the gap between the two figures is a domain effect, and we take the official number as the honest estimate of current utility. Closing it is the most valuable next step: the submit-

ted outputs show that named-entity corruption and news-domain vocabulary, not grammatical competence, account for most of the loss. Further directions include scaling the parallel corpus through community-sourced data and back-translation, applying LoRA (Hu et al., 2021) to adapt the full NLLB-3.3B checkpoint, and running the surrogate-token ablation and multi-seed replication described below.

Limitations

Three limitations bound the claims above. First, the surrogate-token explanation is a hypothesis: we did not decode the same fine-tuned checkpoint under `asm_Beng` and `asm_Latn` on a common test split, so we cannot separate the contribution of the inference-time token from that of fine-tuning itself. The tokeniser behaviour we report (`asm_Latn` → `<unk>`, ID 3) is verified; the causal account of why it helps is not. Second, all reported figures come from a single training run at `seed=42`, so we cannot report variance across seeds. Third, although the model is trained jointly in both directions, only `EN`→`NAG` was evaluated and submitted; we make no claim about `NAG`→`EN` quality. We also note that our internal test set is drawn from the same curated corpus as our training data, so Table 3 measures in-domain performance; the official result in Table 4 is 27.5 BLEU lower and should be preferred as an estimate of open-domain quality.

Reproducibility. Fine-tuned model:
<https://huggingface.co/agnivamaiti/nllb-200-en-nagamese>. Corpus: <https://huggingface.co/datasets/agnivamaiti/nagamese-english-mt>. Training and evaluation scripts: <https://github.com/AgnivaMaiti/NagameseMT>.

References

- Roei Aharoni, Melvin Johnson, and Orhan Firat. 2019. Massively Multilingual Neural Machine Translation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3874–3884. Association for Computational Linguistics.
- Loïc Barrault, Magdalena Biesialska, Ondřej Bojar, Marta R. Costa-jussà, Christian Federmann, Yvette Graham, Barry Haddow, Matthias Huck, Philipp Koehn, and Marcos Zampieri. 2020. Findings of the 2020 Conference on Machine Translation (WMT20). In *Proceedings of the Fifth Conference on Machine Translation*, pages 1–55.
- Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language Models are Few-Shot Learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901.
- Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised Cross-lingual Representation Learning at Scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451.
- Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barraud, Gabriel Mejia Gonzalez, Prangthip Hansanti, John Hoffman, Semarley Jarrett, Kaushik Ram Sadagopan, Dirk Rowe, Shannon Spruit, Chau Tran, Pierre Andrews, Necip Fazil Ayan, Shruti Bhosale, Sergey Edunov, Angela Fan, Cynthia Gao, Vedanuj Goswami, Francisco Guzmán, Philipp Koehn, Alexandre Mourachko, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, and Jeff Wang. 2022. No Language Left Behind: Scaling Human-Centered Machine Translation. *arXiv preprint arXiv:2207.04672*.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1*, pages 4171–4186.
- Abteen Ebrahimi, Manuel Mager, and Katharina Kann. 2022. Low-Resource Machine Translation: A Comprehensive Survey. *arXiv preprint arXiv:2203.04561*.
- Sergey Edunov, Myle Ott, Michael Auli, and David Grangier. 2018. Understanding Back-Translation at Scale. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 489–500.
- Angela Fan, Shruti Bhosale, Holger Schwenk, Zhiyi Ma, Ahmed El-Kishky, Siddharth Goyal, Man-deep Baines, Onur Celebi, Guillaume Wenzek,

- Vishrav Chaudhary, Naman Goyal, Tom Birch, Vitaliy Liptchinsky, Sergey Edunov, Edouard Grave, Michael Auli, and Armand Joulin. 2021. Beyond English-Centric Multilingual Machine Translation. *Journal of Machine Learning Research*, 22(107):1–48.
- Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. LoRA: Low-Rank Adaptation of Large Language Models. In *International Conference on Learning Representations*.
- Melvin Johnson, Mike Schuster, Quoc V. Le, Maxim Krikun, Yonghui Wu, Zhifeng Chen, Nikhil Thorat, Fernanda Viégas, Martin Wattenberg, Greg Corrado, Macduff Hughes, and Jeffrey Dean. 2017. Google’s Multilingual Neural Machine Translation System: Enabling Zero-Shot Translation. *Transactions of the Association for Computational Linguistics*, 5:339–351.
- Nabam Kakum, Sahinur Rahman Laskar, Koj Sambyo, and Partha Pakray. 2023. Neural Machine Translation for Limited Resources English–Nyishi Pair. *Sādhanā*, 48(4):237. Springer.
- Divyanshu Kakwani, Anoop Kunchukuttan, Satish Golla, Gokul N.C., Avik Bhattacharyya, Mitesh M. Khapra, and Pratyush Kumar. 2020. IndicNLP Suite: Monolingual Corpora, Evaluation Benchmarks and Pre-trained Multilingual Language Models for Indian Languages. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4948–4961.
- Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondřej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open Source Toolkit for Statistical Machine Translation. In *Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions*, pages 177–180.
- Guillaume Lample and Alexis Conneau. 2019. Cross-lingual Language Model Pretraining. In *Advances in Neural Information Processing Systems*, volume 32, pages 7059–7069.
- Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880.
- Junjie Hu, Sebastian Ruder, Aditya Siddhant, Graham Neubig, Orhan Firat, and Melvin Johnson. 2020. XTREME: A Massively Multilingual Multi-task Benchmark for Evaluating Cross-lingual Generalisation. In *Proceedings of the 37th International Conference on Machine Learning*, pages 4411–4421.
- Graham Neubig and Junjie Hu. 2018. Rapid Adaptation of Neural Machine Translation to New Languages. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 875–880.
- 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 (WMT)*, pages 654–668.
- 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. 2025a. Findings of WMT 2025 Shared Task on Low-Resource Indic Languages Translation. In *Proceedings of the Tenth Workshop on Machine Translation (WMT)*, pages 532–553, November 8–9.
- 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. 2025b. WMT 2025 Shared Task on Low-Resource Indic Languages Translation: Reference Dataset. LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.
- 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 the WMT 2026 Shared Task on Low-Resource Indic Languages Translation. In *Proceedings of the Eleventh Workshop on Machine Translation (WMT)*, co-located with EMNLP 2026, Budapest, Hungary, 28–29 October.
- 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.
- Maja Popović. 2015. chrF: Character n -gram F-score for Machine Translation Evaluation. In *Proceedings of the Tenth Workshop on Machine Translation*, pages 392–395.
- Matt Post. 2018. A Call for Clarity in Reporting BLEU Scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191.
- Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the Limits of Transfer Learning with a Unified Text-to-Text

- Transformer. *Journal of Machine Learning Research*, 21(140):1–67.
- Gowtham Ramesh, Sumanth Doddapaneni, Aravindh Bheemaraj, Mayank Jobanputra, Raghavan Ak, Ajitesh Sharma, Sujit Sahoo, Harshita Diddee, Divyanshu Kakwani, Navneet Kumar, and others. 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, Marjana Prifti Skenduli, Ravi Shekhar, Mehreen Alam, and Rishemjit Kaur. 2023. Neural Machine Translation for Low-Resource Languages: A Survey. *ACM Computing Surveys*, 55(11):235:1–38.
- 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*, pages 2685–2698.
- Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural Machine Translation of Rare Words with Subword Units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725.
- Matthew Snover, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul. 2006. A Study of Translation Error Rate with Targeted Human Annotation. In *Proceedings of the 7th Conference of the Association for Machine Translation in the Americas*, pages 223–231.
- Yuqing Tang, Chau Tran, Xian Li, Peng-Jen Chen, Naman Goyal, Vishrav Chaudhary, Jiatao Gu, and Angela Fan. 2020. Multilingual Translation with Extensible Multilingual Pretraining and Finetuning. *arXiv preprint arXiv:2008.00401*.
- Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention Is All You Need. In *Advances in Neural Information Processing Systems*, volume 30, pages 5998–6008.
- Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mT5: A Massively Multilingual Pre-trained Text-to-Text Transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498.
- Barret Zoph, Deniz Yuret, Jonathan May, and Kevin Knight. 2016. Transfer Learning for Low-Resource Neural Machine Translation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1568–1575.