

ANVITA : Machine Translation System for Low-resource Indian Languages - Khasi, Nagamese and Tagin

Siva Bhavani J, Daneshwari Kankanwadi[†], Abhinav Gugulothu, Biswajit Paul

{jsbhavani.cair, abhinav.gugulothu, biswajit.cair}@gov.in, [†]daneshwarikankanwadi1599@gmail.com

Centre for Artificial Intelligence and Robotics,
CV Raman Nagar, Bangalore, India

Abstract

In this paper, we present ANVITA machine translation system submitted to WMT 2026 shared task on Low-Resource Indic Language Translation, where the team participated for six translation directions: {Khasi, Nagamese, and Tagin} $\leftrightarrow$ English. Two core ideas shaped the design of ANVITA system. Firstly, transfer learning by symmetric fine-tuning of T5-base model and enhanced cross-lingual transfer from related languages (Assamese to Nagamese) by mitigating script mismatch through transliteration. Secondly, to alleviate data scarcity, primary submissions are trained on datasets augmented with related-language corpora and paraphrased pairs (English sentences), utilizing only the organizers provided datasets. For the contrastive submissions, training corpora are further expanded by compiling synthetic parallel data and related language data followed by corpora distillation with suitable selection strategies. ANVITA submissions are evaluated on the official test sets using multiple metrics which include BLEU, METEOR, TER, CHRF++, BERT score, and COMET. In the primary submissions, ANVITA achieved first rank for English $\rightarrow$ Tagin translation. For the contrastive submissions, ANVITA secured first rank for English $\rightarrow$ Khasi and Khasi $\rightarrow$ English with BLEU scores of 32.89 and 26.48 respectively. Furthermore, the contrastive submissions for Nagamese $\leftrightarrow$ English and Tagin $\rightarrow$ English ranked second, underscoring efficacy of the data augmentation strategies applied.

1 Introduction

India’s linguistic diversity is one of its greatest cultural treasures, yet the lack of comprehensive digital high-quality datasets creates a significant barrier for development of language technology including machine translation (MT). Many major Indian languages are low-resource, particu-

larly the North Eastern languages, such as Khasi, Nagamese, Tagin and lack the necessary data resources, thus posing challenges in building modern MT systems.

To address the scarcity of high-quality training data for the world’s 7,000 languages, the SMOL (Caswell et al., 2025) (Set of Maximal Overall Leverage) open-source suite provides professionally translated datasets at both the sentence and document levels. This initiative aims to bridge the critical performance gap in MT systems for low-resource linguistic contexts. The SEACrowd initiative (Lovenia et al., 2024) establishes a standardized collection of approximately 500 diverse corpora, including numerous parallel datasets to support linguistic resources for nearly 1,000 languages.

Khasi (kha): Khasi, an Austroasiatic language spoken in Northeast India, exhibits significant lexical influence from neighboring Indo-Aryan languages. More than 450 words have been integrated into the Khasi lexicon from Assamese (asm), Bengali, and Hindi, spanning essential semantic domains (Baishya, 2012). Early research on MT resources for Khasi language introduced EnKhCorp1.0 (Laskar et al., 2021), a parallel corpus specifically curated for Khasi $\leftrightarrow$ English machine translation.

Nagamese (nag): Nagamese, an Assamese-lexified creole language with an Indo-Aryan structural base, is built primarily on an Assamese lexical base (Mahanta, 2012) and is spoken in North-east India. Nagamese shares a vast percentage of its lexicon and roots with Assamese. This paper (Maiti et al., 2025), released a human-validated 10K pair conversational dataset and NagaNLP toolkit¹. This toolkit serve as a foundational resource for downstream NLP tasks for Nagamese. The "NagaKotha" system, introduced by (Yadav

¹<https://huggingface.co/agnivamaiti>

et al., 2024), aims to build a specialized translation system to overcome the challenges of translating Nagamese language.

Tagin (tgj): Tagin, a Sino-Tibetan language spoken in Northeast India is a UNESCO classified "definitely endangered" language. Tagin shares the closest structural, phonological, and lexical similarities with Nyishi language, again a low-resource language. GinLish corpus (Dugi, 2024) is a parallel corpus of English and Tagin, developed to facilitate research in machine translation, linguistic analysis, and related fields.

NE-BERT (Nyalang, 2026) is a domain-specific 149M parameter multilingual language model trained on 8.3M sentences spanning 11 Northeastern languages, designed to enhance NLP capabilities of nine highly underrepresented Northeast Indian languages: Meitei, Assamese, Khasi, Mizo, Nyishi, Garo, Naga, Kokborok, and Pnar. Kren-M (Labs, 2025) serves as a primary technological pillar within Northeast India's ongoing efforts for AI-driven language preservation. This work (Mitri et al., 2025) leverages the zero-shot and fine-tuning capabilities of generative Large Language Models (LLMs) specifically for Khasi abstractive text summarization.

Google T5 model (Raffel et al., 2020) was primarily trained on massive text corpora of English data (C4 corpus), along with some baseline exposure to French, German, and Romanian from its training mixture. T5-base is well-suited for machine translation (Usui and Komiya, 2023) due to its encoder-decoder architecture and text-to-text paradigm. However, achieving enhanced translation performance requires fine-tuning on target language and domain-specific parallel corpora. T5-base is employed as the foundational pre-trained model for majority of the ANVITA primary and contrastive submissions. In this work, ANVITA explores how transfer learning can be effectively utilized along with different data augmentation strategies to overcome data scarcity and improve machine translation quality for the low-resource Indian languages.

2 Task Setup

WMT 2026 Low-Resource Indic Language Translation shared task continued its focus on creation of MT models that perform robustly, despite limited parallel data availability by integrating innovative training techniques and data augmentation

strategies into the MT model building pipelines. The participating languages are classified into two categories based on the volume of training corpora made available by the organizers. Category 1, designated as the moderate—resource tier, encompassed six languages and twelve translation directions — Assamese, Mizo, Khasi, Manipuri, Meiti, and Nyishi paired with English. In contrast, Category 2 represented the low—resource tier, consisting of five languages and ten translation directions — Bodo, Kokborok, Karbi, Nagamese and Tagin paired with English. Translation submissions are capped at three per language pair, consisting of one constrained Primary system utilizing only the organizer's provided resources, monolingual resources, and public pre-trained models; and two unrestricted Contrastive systems, which permit usage of auxiliary external data. Comprehensive details of the task is available in the findings of WMT 2026 shared task on Low-Resource Indic Language Translation (Pakray et al., 2026).

The WMT Indic MT shared task utilized blind test sets, with the suite of evaluation metrics (BLEU, METEOR, TER, chrF++, BERTScore, and COMET) designated by the task organizers. Furthermore, the organizers independently executed the evaluation pipeline by scoring the submitted translations against hidden reference sets and subsequently publishing the official evaluation results on WMT website².

ANVITA made fifteen submissions (six Primary and nine Contrastive) addressing six translation directions: Khasi ↔ English, Nagamese ↔ English, and Tagin ↔ English.

3 Related Work

While Neural Machine Translation (NMT) has advanced significantly, building robust MT systems with limited training corpora remains a complex and vital objective. Recent studies have explored balancing translation performance with computational efficiency and fine-tuning strategies to adapt pre-trained language and machine translation models for the under-represented languages; for instance robust models like mBART, mT5, T5, IndicTrans2 or NLLB (Galiano-Jiménez et al., 2023; Sant et al., 2024; Sahoo et al., 2024; Wei et al., 2024) successfully adapted for low resource Neural Machine Translation with varying success rates. However, LLMs are yet to catch up (Pucin-

²<https://www2.statmt.org/wmt26/indic-mt-task.html>

Language Pair	Language Family, Script	Data(Org) (#sent)	Data(PP) (#sent)	Total words	Unique words	Average sentence length (# words)
Khasi(kha)-English(en)	Austroasiatic Latin	26,000	25,995	(kha) 966,353 (en) 798,291	(kha) 8,123 (en) 12,778	(kha) 37.16 (en) 30.7
Nagamese(nag)-English(en)	Creole Latin	2,000	1,997	(nag) 27,747 (en) 26,480	(nag) 5,490 (en) 5,592	(nag) 13.89 (en) 13.24
Tagin(tgj)-English(en)	Sino-Tibetan Latin	5,343	5,108	(tgj) 26,072 (en) 26,575	(tgj) 5,161 (en) 3,477	(tgj) 4.88 (en) 7.97
Assamese-English	Indo-Aryan Bengali-Assamese	54,000	53,923	(asm) 787,788 (en) 901,984	(asm) 62,826 (en) 31,379	(asm) 14.59 (en) 16.7

Data(Org): Parallel data before data processing

Data(PP): Parallel data after data processing

Table 1: Statistics of WMT 2026 official training corpora (provided by the organizers) and preprocessing results

Language Pairs	Data(PP) (#sent)	Contrastive I	Contrastive II
kha-en	8,018,784	Yes	Yes
nag-en	12,948	Yes	Yes
tgj-en	1,981,427	Yes	N/A

Table 2: Statistics of synthetic parallel corpora compiled for ANVITA contrastive systems

Translation direction	Data (#sent)	Total words	Unique words	Average sentence length (#words)
en → kha	1,000	11,903	3,654	11.9
kha → en	1,000	22,699	2,313	22.7
en → nag	1,000	11,903	3,654	11.9
nag → en	1,000	9,933	2,301	9.93
en → tgj	1,000	11,903	3,654	11.9
tgj → en	1,000	12,339	2,418	12.34

Table 3: Statistics of respective source languages of WMT 2026 official Test set

Domain	Confidence Threshold	No. of sentence
Business	≥ 50	107,092
Culture	≥ 50	361,049
Education	≥ 90	10,905
Entertainment	≥ 80	48,632
Environment	≥ 80	83,069
Finance	≥ 90	7,089
Food and Cooking	≥ 90	10,370
Geography	≥ 90	3,653
Healthcare	≥ 70	37,441
Legal	≥ 80	29,210
Medical	≥ 40	47,595
Politics	≥ 90	87,752
Science	≥ 90	20,313
Sports	≥ 70	153,724
Technology	≥ 80	51,471
Travel and Tourism	≥ 70	11,984

Table 4: Statistics of domain classification of synthetic Khasi-English corpora (1M+)

skaitė and Mitkov, 2025) and the recent studies introduced various approaches (Gain et al., 2026; Nag et al., 2025) to improve low-resource translation performance.

4 Dataset

WMT 2026 official training corpora provided by the organizers is outlined in Table 1, along with corpus size variability which ranges between 2K to 26K sentences w.r.t Khasi, Nagamese and Tagin languages. ANVITA primary submissions are solely based on the official parallel datasets provided by the organizers.

Synthetic Corpora for Primary Systems :

To generate additional sentence pairs utilizing the primary data, English sentences are paraphrased and paired with the respective unchanged

Khasi, Nagamese and Tagin sentences. For paraphrasing of English sentences, round trip translation strategy is used where English sentences are first translated to intermediate non-English languages and subsequently translated back to English. Open source machine translation models of high resource languages to/from English with good BLEU scores are employed for both forward and backward translations. This iterative pipeline leveraged multiple open translation models³ to generate a diverse set of paraphrased English sentences. Further, quality of these paraphrased sentences are assessed using sentence similarity tool⁴

³<https://huggingface.co/Helsinki-NLP>

⁴<https://huggingface.co/sentence->

and included without filtering.

Additional Corpora for Contrastive Systems:

Low-resource language translation requires diverse and quality parallel data to bridge the linguistic gaps between widely represented English and low-resource Indic languages. To address this, ANVITA leveraged curated bilingual resources to enhance performance and generalization capabilities of the machine translation models. The additional parallel corpora curated for the contrastive submissions is summarized in Table 2. Our data collection strategy for Contrastive systems include harvesting of monolingual and parallel corpora from web and creation of synthetic parallel corpora using back-translation, followed by corpora distillation, primarily using round-trip translation similarity.

Khasi: Synthetic English–Khasi parallel corpora of size 8M have been curated from three sources (a) Kaggle ⁵, (b) Github ⁶ and (c) Monolingual Khasi sentences harvested from news portals⁷ and back translated to English.

Nagamese: For compiling Nagamese-English parallel corpora of 12.9K, first FNPC-25 corpus⁸ comprising parallel sentence pairs majorly sourced from Bible and conversational texts are included. Further, to ensure broader linguistic generalization, ANVITA utilized 3.3K Nagamese-English machine translation corpus hosted on Hugging Face⁹ which covers diverse domains including sports, narration, and dialogue to ensure models generalize beyond a single, narrow domain. Together, these datasets provides reasonable corpora for fine-tuning pre-trained translation models.

Tagin: A synthetic Tagin-English parallel corpora of 1.98M sentence pairs are created as follows: 21M English sentences are collected from Samantar¹⁰, Tatoeba ¹¹, and TinyStories ¹². Further forward and backward translations are generated using a public Tagin ↔ English

transformers/paraphrase-MiniLM-L6-v2

⁵<https://www.kaggle.com/datasets/user10383/english-khasi-synthetic-parallel-corpus>

⁶<https://github.com/kirbynuggets/t5-kha-eng>

⁷<https://www.urupang.com/>, <https://unongsainhima.com/>

⁸<https://www.kaggle.com/datasets/belovedorange/first-nagamese-english-parallel-corpus-fnpc-25>

⁹<https://huggingface.co/datasets/agnivamaiti/nagamese-english-mt>

¹⁰<https://opus.nlpl.eu/datasets/Samanantar?pair=en&hi>

¹¹<https://tatoeba.org/en/downloads>

¹²<https://huggingface.co/datasets/roneneldan/TinyStories>

translation model¹³ and subsequently 1.9M Tagin-English sentence pairs are filtered out based on round-trip translation similarity scores, computed using paraphrase-MiniLM-L6-v2¹⁴ sentence similarity assessment tool.

Validation Set The data provided by the organizers is partitioned into training and validation sets. Specifically, for the Tagin-English and Nagamese-English language pairs, the validation set is formed by reserving 150 sentences, while the remainder of the data is allocated for training. For the Khasi-English language pairs, the previous year’s WMT 2025 test set was adopted as the validation set for all experiments.

Test Set The WMT Indic MT shared task uses blind test sets, where the ground-truth references remain withheld from participating teams. The official test set utilized by the organizer for performance evaluation is outlined in Table 3.

5 ANVITA Machine Translation System

This section describes design of ANVITA system for three low resource languages Khasi, Nagamese and Tagin.

5.1 Data Preprocessing

The preprocessing pipeline involved cleaning the corpora listed in Tables 1 and Table 2 using language agnostic noise reduction strategies established in (Vegi et al., 2021, 2022), a step intended to improve the overall quality of the training data. The original dataset statistics are presented in Table 1 and preprocessing outcomes are presented in Table 1. Further for Indic transliterations wherever applied has used IndicXlit tool¹⁵ and paraphrasing wherever mentioned has used technique mentioned in Paraphrasing for Augmenting Primary Dataset⁴.

5.2 Training Methodology With T5-base:

This section details the training methodology employed using T5-base architecture. The considerations for choice of T5’s over massive multilingual models like mT5, mBART, NLLB, IndicTrans2 are the compact, Latin-only vocabulary, which potentially enhances script-level transfer with a reduced risk of language bleeding, as

¹³<https://huggingface.co/repleeka/ginlishMT-mbart50-tgj-en-bi>

¹⁴<https://huggingface.co/sentence-transformers/paraphrase-MiniLM-L6-v2>

¹⁵<https://github.com/AI4Bharat/IndicXlit>

Khasi, Nagamese, and Tagin all utilize the Latin script. However, an empirical validation of this architectural choice remains to be explicitly conducted. The details regarding experimental procedure and datasets utilized are captured in the subsequent subsections.

5.2.1 Training Data Preparation

To facilitate bidirectional translation for symmetric fine-tuning, the dataset was processed by prepending specific language-pair prefixes to the source sentences. This preprocessed data was used as input for training.

For the Indic-to-English direction, input sequences were prefixed with "translate [Indic Language] to English: ", while for the English-to-Indic direction, the prefix "translate English to [Indic Language]: " was used. Consequently, for each parallel pair for instance like (*kha*, *en*), two training samples were generated: one mapping *kha* $\rightarrow$ *en* and another mapping *en* $\rightarrow$ *kha*. This approach effectively doubles the training corpus size, enables deep alignment and better generalization. Following the addition of these prefixes, the entire dataset was shuffled to ensure there are no order biases during training.

5.2.2 Sub-word Tokenization and Language Tagging of Sentences

The pre-trained T5 tokenizer was used without extending its original vocabulary with custom sub-word tokens. Instead, language-specific prefixes were employed to identify the source and target languages for both translation directions. For instance, prefixes containing the relevant language identifier (e.g., Khasi, Tagin, or Nagamese) were prepended to the input sentences to facilitate direction-specific translation learning.

5.2.3 Batch Preparation for Symmetric Fine-Tuning

Batches were constructed using default dataloader from Hugging Face toolkit and the Hugging Face DataCollatorForSeq2Seq¹⁶ was used to handle sequence padding. As the dataset was shuffled prior to training, each batch comprised a heterogeneous distribution of samples covering both translation directions. This batch composition enable that the cross-entropy loss reflected the model's performance on both Indic-to-English and English-

to-Indic translation objectives within each training step.

5.2.4 Hyperparameter Settings and Evaluation During Training

In the primary experimental setting, models were trained for 12–15 epochs with an effective batch size of 128, distributed across multiple GPUs. In the contrastive setting, training was extended to 20 epochs with an effective batch size of 256. Throughout the training process, the models were evaluated on automated metrics at the end of each epoch to monitor performance.

5.2.5 Loss Function

A cross-entropy loss function with label smoothing was employed for symmetric fine-tuning, where each training instance represents a single translation direction. Consequently, the batch level loss comprised the aggregated losses from both translation directions. These losses were averaged across the batch to facilitate gradient-based weight optimization.

All translation systems, encompassing both primary and contrastive models, were trained using the identical methodology, with the exception of Nagamese $\rightarrow$ English primary submission.

5.3 Nagamese $\leftrightarrow$ English Translation Systems

For **Nagamese $\rightarrow$ English (primary)** translation system, Assamese was employed as a linguistic bridge, considering the deep lexical and structural overlapping characteristics of Nagamese as an Assamese-lexified creole. To formally define the pipeline proceeds through two main phases: transliteration followed by translation.

Transliteration: Nagamese text was converted into the Bengali script (the same script as used in organizers provided Assamese-English parallel dataset) for better cross-lingual transfer, utilizing IndicXlit tool aligning with the script supported by IndicTrans2 (Gala et al., 2023), a open source multilingual translation model covering 22 official Indian languages including Assamese and excluding Nagamese.

Translation: Zero-shot translation was performed via the IndicTrans2 model on the Nagamese to Assamese transliterated text to generate English translations.

For the **English $\rightarrow$ Nagamese (primary) system**, English-Assamese dataset provided by the

¹⁶https://huggingface.co/docs/transformers/en/main_classes/data_collator

Optimizer	AdamW
learning rate	1e-4
weight decay by optimiser	0.01
warmup steps	500
learning rate scheduler	linear with warmup
precision	fp32
maximum number of epochs (Primary /contrastive setting)	15 /20
metric for best model	BLEU

Table 5: Parameters used for fine-tuning T5-base

organizers was used as an auxiliary data resource. The English-Assamese data was transliterated into Assamese Latin script and augmented with the English-Nagamese dataset to create a total training corpus of 56K sentences, followed by model training as described in subsection 5.2.

To construct the **Nagamese ↔ English (contrastive)** translation systems, related-language parallel corpora from Samanantar (Ramesh et al., 2022), specifically Assamese–English, Bengali–English and Hindi–English language pairs were augmented with the Nagamese–English instances to compile the training data. To facilitate cross-lingual knowledge transfer, these related language corpora were incorporated into the training datasets at varying mixing proportions for Contrastive Systems I and II. For Contrastive System I, the training corpus comprised 12K Nagamese–English, 900K Assamese–English, 300K Bengali–English, and 50K Hindi–English parallel sentence pairs. For Contrastive System II, the training data consisted of 12K Nagamese–English, 100K Assamese–English, 75K Bengali–English, and 50K Hindi–English instances. Further, Assamese, Bengali, and Hindi texts were transliterated to Latin script for better cross-lingual transfer. Model training followed is as described in subsection 5.2.

5.4 Tagin ↔ English Translation Systems

For **Tagin ↔ English (primary)** system, training was carried out as described in the subsection 5.2. The training data was curated by combining related language data along with paraphrased sentences, all from the organizers data and includes (i) Tagin-English and Nyishi-English organizers data (ii) Assamese-English organizer data transliterated to Assamese Latin script using IndicXlit tool while keeping the English text intact and (iii) Created 20K+ paraphrased Tagin-English sen-

tence pairs by paraphrasing every Tagin-English organizers data into 4 variations and by keeping the Tagin text intact.

The **Tagin ↔ English (contrastive)** system was developed using additional 1.9M synthetic parallel corpora as described in section 4.

5.5 Khasi ↔ English Translation Systems

For primary and contrastive Khasi ↔ English translation systems, training was carried out as described in the subsection 5.2.

The **Khasi ↔ English (primary)** submissions also used related language data along with paraphrased sentences, all from the organizers data and includes (i) Khasi-English dataset provided by organizers (ii) Created paraphrased 50K+ sentence pairs from Khasi-English organizers data by paraphrasing every English sentence into 2 variations and by keeping the Khasi text intact (iii) Assamese-English organizers data transliterated to Assamese Latin script while keeping the English text intact.

For **Khasi ↔ English (contrastive)** submission, datasets used are as presented in the Table 2. The contrastive systems were developed using distinct data sampling strategies. Contrastive System II was trained on a randomly selected subset of 4M instances from the total compiled corpus of 8M. Conversely, contrastive system I utilized a combined dataset of 1M randomly sampled instances and 1M instances selected based on pre-defined information domains (Table 4). For selecting 1M domain-classified instances, classification was performed on the remaining 7M sentences using bart-large¹⁷ model, with a confidence threshold as captured in Table 4. To align the training data with the target evaluation domains, domain distribution of the previous year’s WMT 2025 Khasi-English test set was considered as a reference and 1 million sentences matching WMT 2025 test set domain distribution were compiled. Furthermore, confidence thresholds were dynamically adjusted to filter the data and ensure adherence to this reference domain distribution.

5.6 Model Training and Experiment Details

All the experiments were conducted on NVIDIA DGX with 8xA100-80GB and HPE Cray with 4xH100-80GB GPU systems. For {Tagin, Khasi} ↔ English primary setting, the effective batch-

¹⁷<https://huggingface.co/facebook/bart-large-mnli>

Lang. Direction [# sent]	Primary/ Contrastive	BLEU	METEOR	TER	chrF++	BERTScore	COMET	Rank
Khasi (kha) → English (en) [1000]	primary contrastive	5.17 26.48	22.98 55.26	119.93 62.48	25.78 52.31	88.01 94.55	50.7 75.62	Sixth First
English (en) → Khasi (kha) [1000]	primary contrastive-2 contrastive-1	10.08 32.89 32.46	28.94 56.75 57.5	77.15 48.53 48.65	30.46 54.84 55.54	74.13 84.51 84.66	57.43 69.69 69.76	Fifth First Second
Nagamese (nag) → English (en) [1000]	primary contrastive-2 contrastive-1	10.91 13.82 14.72	37.91 39.43 40.78	80.09 73.78 72.56	36.1 39.37 40.48	90.96 92.52 92.7	66.47 72.95 73.57	Third Second Third
English (en) → Nagamese (nag) [1000]	primary contrastive-2 contrastive-1	16.37 16.14 15.88	52.05 50.13 50.25	81.65 82.28 83.73	51.81 50.22 50.48	80.25 79.63 79.66	65.16 64.14 63.81	Third Second Third
Tagin (tgj) → English (en) [1000]	primary contrastive	2.43 3.52	12.59 15.46	100.71 91.48	16.72 17.96	86.52 87.1	47.01 49.56	Fourth Second
English (en) → Tagin (tgj) [1000]	primary contrastive	4.81 3.71	15.04 20.12	92.58 99.11	22.39 28	70.49 72.76	52.61 58.65	First Fourth

Table 6: Performance of ANVITA on the WMT 2026 official Test set

size was set to 128 (by using 4 GPUs and 2 gradient accumulation steps) during training. For {Nagamese, Khasi, Tagin} ↔ English contrastive setting, effective batch-size of 256 (by using 4 GPUs and 4 gradient accumulation steps) was used during training. Gradient accumulation was used to avoid out-of-memory issues while training. Other training parameters for all the experiments are shown in Table 5.

6 Evaluation and Result Analysis

To evaluate the submitted primary and contrastive systems, the WMT organizers used blind test set (Table 3) and suite of select metrics (BLEU, METEOR, TER, chrF++, BERTScore, and COMET). Furthermore, the organizers independently executed the evaluation pipeline by scoring the submitted translations against the hidden reference sets. The performance of ANVITA systems on the official test, as released by the WMT organizers are captured in Table 6. Further, findings of the WMT 2026 shared task on Low-Resource Indic Language Translation (Pakray et al., 2026) provides a comprehensive evaluation of the participating systems, benchmark results, and performance trends to highlight the current capabilities.

Primary Systems: As shown in Table 6 using 2K, 26K, and 5K+ parallel corpora, ANVITA MT systems for English to {Nagamese, Khasi, Tagin} attained BLEU scores of 16.37, 10.08, 4.81 respectively. For {Nagamese, Khasi, and Tagin} to English direction BLEU scores of 10.91, 5.17, 2.43 respectively were attained. The relatively higher score for Nagamese with just 2K corpora

is due to its significant similarity with Assamese, a relatively more resourced language. The lower scores for Khasi and Tagin may be attributed to the following: (i) Lack of these language support in T5-base leading to a weaker transfer and (ii) Linguistic complexity of these languages like intricate agglutinative structures and word order patterns that differ substantially from English, complicating the learning process for the model with the existing very low corpora.

Contrastive Systems: Performance of ANVITA on the WMT 2026 test set for English to {Nagamese, Khasi, Tagin} attained BLEU scores of 16.14, 32.89, 3.71 and 14.72, 26.48, 3.52 for {Nagamese, Khasi, and Tagin} to English respectively. For Khasi ↔ English, the contrastive systems showed significant performance enhancement upon incorporating the additional curated dataset. The English → Nagamese contrastive systems however showed no significant variation between the 2K and 12K+ training data configurations. Similar trend is also noticed with Tagin ↔ English systems. The lack of significant improvement in BLEU scores could be that the additional data lacks high-quality alignments or domain relevance. Furthermore, the average sentence length in the WMT 2026 Tagin test set is 12.34 words, compared to just 4.88 and 8.29 words for the organizers provided Tagin training dataset and the Tagin contrastive datasets respectively. This difference in sentence length is likely one of the contributing factors for the lower BLEU scores observed on the WMT 2026 test sets.

Overall, finetuned T5-base also exhibited effective

absorption of syntactic and semantic patterns in case of contrastive {Khasi, Nagamese} $\leftrightarrow$ English MT systems.

7 Conclusions

ANVITA submissions have demonstrated encouraging MT performance for the three low-resource Indian languages by employing effective training strategies utilizing transfer learning along with data augmentation, including utilization of synthetic parallel corpora and related language corpora with distillation for mitigating the limitation of training data. Despite substantial data scarcity barrier, ANVITA primary submissions achieved BLEU scores between 2.43 - 10.91, utilizing organizers datasets ranging only 2K to 26K for Nagamese, Tagin, and Khasi and the submitted English $\rightarrow$ Tagin system secured First rank. By employing robust data augmentation and refinement techniques, ANVITA contrastive systems secured first and second ranks for Khasi $\leftrightarrow$ English and Nagamese $\leftrightarrow$ English translations respectively. The contrastive submissions, trained on datasets varying from 14K to 4M samples, yielded competitive performance with BLEU scores spanning 10.08–32.89. The finetuned T5-base (220M) struck a balance between computational efficiency and translation quality with novel training methodology.

Despite the utilization of data augmentation and effective training techniques, the performance metrics underscores a critical requirement of improvement in strategies for a robust MT system.

8 Future Work

Future work should focus on assessing a few other potential pre-trained models for enhanced transfer learning with maximal representational overlaps and further fine-tuning with the aligned objectives and suitable loss functions. Further in depth ablation studies covering some of the techniques used in ANVITA systems are also likely to add value in the journey of low resource machine translation.

9 Contributions

Siva Bhavani J and Biswajit Paul conceptualized and designed ANVITA system. Abhinav Gugu-lothu led the compilation and distillation of the corpora, preparation of the training data, and evaluation. Daneshwari Kankanwadi and Siva Bhavani conducted the model training and core experi-

ments. The manuscript was prepared by Siva Bhavani J.

Acknowledgments

The authors would like to thank Director, CAIR for his constant encouragement and support. The authors would also like to thank Prasanna Kumar KR, Chitra Viswanathan, and N. Venkataramanan for their guidance and enablement.

References

- A. K. Baishya. 2012. Indo-aryan loan words in khasi: An introductory note. volume 3, pages 55–60. Global Media Journal, Indian Edition.,
- Isaac Caswell, Elizabeth Nielsen, Jiaming Luo, Colin Cherry, Geza Kovacs, Hadar Shemtov, Partha Talukdar, Dinesh Tewari, Baba Mamadi Diane, Djibrila Diane, Solo Farabado Cissé, Koulako Moussa Doumbouya, Edoardo Ferrante, Alessandro Giasoni, Christopher Homan, Mamadou K. Keita, Sudhamoy DebBarma, Ali Kuzhuget, David Anugraha, Muhammad Ravi Shulthan Habibi, Sina Ahmadi, Anthony Munthali, Jonathan Mingfei Liu, and Jonathan Eng. 2025. [SMOL: Professionally translated parallel data for 115 under-represented languages](#). In *Proceedings of the Tenth Conference on Machine Translation*, pages 1103–1123, Suzhou, China. Association for Computational Linguistics.
- Tungon Dugi. 2024. [Ginlish v0.1.0: First ever tagin and english parallel corpus \[data set\]](#). NIT Arunachal Pradesh.
- Baban Gain, Dibyanayan Bandyopadhyay, Asif Ekbal, and Trilok Nath Singh. 2026. Bridging the linguistic divide: A survey on leveraging large language models for machine translation. *Language Resources and Evaluation*, 60(2):40.
- Jay Gala, Pranjal A Chitale, A K Raghavan, Varun Gumma, Sumanth Doddapaneni, Aswanth Kumar M, Janki Atul Nawale, Anupama Sujatha, Ratish Puduppully, 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*.
- Aarón Galiano-Jiménez, Felipe Sánchez-Martínez, Víctor M. Sánchez-Cartagena, and Juan Antonio Pérez-Ortiz. 2023. [Exploiting large pre-trained models for low-resource neural machine translation](#). In *Proceedings of the 24th Annual Conference of the European Association for Machine Translation*, pages 59–68, Tampere, Finland. European Association for Machine Translation.
- MWire Labs. 2025. [Kren-m: A bilingual language model for khasi](#).

- Sahinur Rahman Laskar, Abdullah Faiz Ur Rahman Khilji Darsh Kaushik, Partha Pakray, and Sivaji Bandyopadhyay. 2021. [EnKhCorp1.0: An English–Khasi corpus](#). In *Proceedings of the 4th Workshop on Technologies for MT of Low Resource Languages (LoResMT2021)*, pages 89–95, Virtual. Association for Machine Translation in the Americas.
- Holy Lovenia, Rahmad Mahendra, Salsabil Maulana Akbar, Lester James V. Miranda, Jennifer Santoso, Elyanah Aco, Akhdan Fadhilah, Jonibek Mansurov, Joseph Marvin Imperial, Onno P. Kampman, Joel Ruben Antony Moniz, Muhammad Ravi Shulthan Habibi, Frederikus Hudi, Railey Montalan, Ryan Ignatius, Joanito Agili Lopo, William Nixon, Börje F. Karlsson, James Jaya, Ryandito Diandaru, Yuze Gao, Patrick Amadeus, Bin Wang, Jan Christian Blaise Cruz, Chenxi Whitehouse, Ivan Halim Parmonangan, Maria Khelli, Wenyu Zhang, Lucky Susanto, Reynard Adha Ryanda, Sonny Lazuardi Hermawan, Dan John Velasco, Muhammad Dehan Al Kautsar, Willy Fitra Hendria, Yasmin Moslem, Noah Flynn, Muhammad Farid Adilazuarda, Haochen Li, Johannes Lee, R. Damanhuri, Shuo Sun, Muhammad Reza Qorib, Amirbek Djanibekov, Wei Qi Leong, Quyet V. Do, Niklas Muennighoff, Tanrada Pansuwan, Ilham Firdausi Putra, Yan Xu, Tai Ngee Chia, Ayu Purwarianti, Sebastian Ruder, William Tjhi, Peerat Limkonchotiwat, Alham Fikri Aji, Sedrick Keh, Genta Indra Winata, Ruochen Zhang, Fajri Koto, Zheng-Xin Yong, and Samuel Cahyawijaya. 2024. [SEACrowd: A multilingual multimodal data hub and benchmark suite for Southeast Asian languages](#). In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 5155–5203, Miami, Florida, USA. Association for Computational Linguistics.
- Shakuntala Mahanta. 2012. [Assamese](#). *Journal of the International Phonetic Association*, 42(2):217–224.
- Agniva Maiti, Manya Pandey, and Murari Mandal. 2025. [Naganlp: Bootstrapping nlp for low-resource nagamese creole with human-in-the-loop synthetic data](#).
- Aiom Minnette Mitri, Goutam Saha, Saralin A Lyngdoh, and Arnab Kumar Maji. 2025. [Abstractive summarization of khasi texts using pretrained large language models](#). *Procedia Computer Science*, 258:4117–4127. International Conference on Machine Learning and Data Engineering.
- Arijit Nag, Soumen Chakrabarti, Animesh Mukherjee, and Niloy Ganguly. 2025. [Efficient continual pre-training of LLMs for low-resource languages](#). In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 3: Industry Track)*, pages 304–317, Albuquerque, New Mexico. Association for Computational Linguistics.
- Badal Nyalang. 2026. [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.
- 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 Wajri, 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 Workshop on Machine Translation (WMT) under EMNLP 2026*, Budapest, Hungary.
- Julita JP Pucinskaite and Ruslan Mitkov. 2025. [Evaluating the LLM and NMT models in translating low-resourced languages](#). In *Proceedings of the First Workshop on Comparative Performance Evaluation: From Rules to Language Models*, pages 123–133, Varna, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.
- 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, Mahalakshmi J, Divyanshu Kakwani, Navneet Kumar, Aswin Pradeep, Srihari Nagaraj, Kumar Deepak, Vivek Raghavan, Anoop Kunchukuttan, Pratyush Kumar, and Mitesh Shantadevi Khapra. 2022. [Samanantar: The largest publicly available parallel corpora collection for 11 Indic languages](#). *Transactions of the Association for Computational Linguistics*, 10:145–162.
- Pramit Sahoo, Maharaj Brahma, and Maunendra Sankar Desarkar. 2024. [NLIP_Lab-IITH low-resource MT system for WMT24 Indic MT shared task](#). In *Proceedings of the Ninth Conference on Machine Translation*, pages 781–787, Miami, Florida, USA. Association for Computational Linguistics.
- Aleix Sant, Daniel Bardanca, José Ramon Pichel Campos, Francesca De Luca Fornaciari, Carlos Escolano, Javier Garcia Gilabert, Pablo Gamallo, Audrey Mash, Xixian Liao, and Maite Melero. 2024. [Training and fine-tuning NMT models for low-resource languages using apertium-based synthetic corpora](#). In *Proceedings of the Ninth Conference on Machine Translation*, pages 925–933, Miami, Florida, USA. Association for Computational Linguistics.
- Hisao Usui and Kanako Komiya. 2023. [Translation from historical to contemporary Japanese using Japanese t5](#). In *Proceedings of the Joint 3rd International Conference on Natural Language Processing*

- for Digital Humanities and 8th International Workshop on Computational Linguistics for Uralic Languages*, pages 27–35, Tokyo, Japan. Association for Computational Linguistics.
- Pavanpankaj Vegi, Sivabhavani J, Biswajit Paul, Prasanna K R, and Chitra Viswanathan. 2022. [ANVITA-African: A multilingual neural machine translation system for African languages](#). In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 1090–1097, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.
- Pavanpankaj Vegi, Sivabhavani J, Biswajit Paul, Chitra Viswanathan, and Prasanna Kumar K R. 2021. [ANVITA machine translation system for WAT 2021 MultiIndicMT shared task](#). In *Proceedings of the 8th Workshop on Asian Translation (WAT2021)*, pages 244–249, Online. Association for Computational Linguistics.
- Bin Wei, Zheng Jiawei, Zongyao Li, Zhanglin Wu, Jiaxin Guo, Daimeng Wei, Zhiqiang Rao, Shaojun Li, Yuanchang Luo, Hengchao Shang, Jinlong Yang, Yuhao Xie, and Hao Yang. 2024. [Machine translation advancements of low-resource Indian languages by transfer learning](#). In *Proceedings of the Ninth Conference on Machine Translation*, pages 775–780, Miami, Florida, USA. Association for Computational Linguistics.
- Satyapal Yadav, Prakash Marimuthu, and Shakila Baskaran. 2024. [Nagakotha: A text-based nagamese \(indian tribal language\) to english translator](#). *2024 International Conference on Advancement in Renewable Energy and Intelligent Systems (AREIS)*, pages 1–5.