Findings of the MMLoSo 2025 Shared Task on Machine Translation into Tribal Languages

Pooja Singh[1], **Sandeep Chatterjee**[2], **Gullal S. Cheema**[3], **Amrit Singh Bedi**[4], **Tanmoy Chakraborty**[1], **Sandeep Kumar**[1], **Ankita Shukla**[5]
[1]IIT-Delhi, India [2]ISI, India [3]Leibniz Uni. Hannover, Germany
[4]Univ. of Central Florida, USA [5]Univ. of Nevada, Reno, USA
eez228470@iitd.ac.in, cs2318@isical.ac.in, gullal.cheema@stud.uni-hannover.de,
amritbedi@ucf.edu, tanchak@iitd.ac.in, ksanpdeep@iitd.ac.in, ankitas@unr.edu

Abstract

This paper presents the findings of the MMLoSo Shared Task on Machine Translation. The competition features four tribal languages from India: Bhili, Mundari, Gondi, and Santali, each with 20,000 high-quality parallel sentence pairs and a 16,000-sentence evaluation set. A total of 18 teams submitted across all language pairs. The shared task addresses the challenges of translating India's severely low-resource tribal languages, which, despite having millions of speakers, remain digitally marginalized due to limited textual resources, diverse scripts, rich morphology, and minimal publicly available parallel corpora. Systems were ranked using a weighted composite score combining BLEU (60%) and chrF (40%) to balance structural accuracy and character-level fluency. The best-performing system leveraged IndicTrans2 with directional LoRA adapters and reverse-model reranking. This work establishes the first reproducible benchmark for machine translation in these tribal languages. All datasets, baseline models, and system outputs are publicly released to support continued research in India's tribal language technologies.

1 Introduction

India is home to an extraordinary diversity of languages, including more than 460 tribal languages documented in the 2011 Census. Many of these languages have substantial speaker populations, yet they remain severely underrepresented in modern NLP research. The primary reasons include limited digital presence, inconsistent or evolving orthographies, and the absence of large, standardized parallel corpora (Joshi et al., 2020; Nekoto et al., 2020). As a result, the rapid progress in deep learning and multilingual LLMs has had little impact on the communities that speak these languages.

The challenges faced by tribal communities extend beyond language processing. Data scarcity affects crucial sectors such as healthcare, education, biodiversity monitoring, and governance. Practitioners working in these areas often operate in environments with limited connectivity and minimal technical infrastructure. For AI systems to be useful in such contexts, they must be resilient to missing modalities, noisy inputs, and shifting distributions (Sun et al., 2023). This shared task therefore focuses not only on translation accuracy but also on robustness under real-world constraints.

The MMLoSo Language Challenge 2025 (Shukla et al., 2025) aims to directly address this gap by advancing research in machine translation for India's tribal and very low-resource languages. While recent large-scale multilingual and multimodal models (Conneau et al., 2020; Alayrac et al., 2022; Li et al., 2023) have demonstrated impressive generalization abilities, they are primarily trained on high-resource languages and domains. Tribal languages, which are often oral, regionally grounded, and culturally specific, fall outside the distribution of mainstream training datasets. This disconnect results in poor performance, hallucination, and limited utility for real-world applications.

2 Task Description

This shared task focuses on developing neural machine translation systems for bidirectional translation between high-resource languages (Hindi, English) and four low-resource Indian tribal languages: Bhili, Mundari, Gondi, and Santali. The primary challenge lies in building effective models despite limited parallel data, diverse writing

Proceedings of MMLoSo 2025: First Workshop on Multimodal Models for Low-Resource Contexts and Social Impact, pages 121–129
December 23, 2025 ©2025 Association for Computational Linguistics

systems, and domain-specific vocabulary characteristic of tribal communities.

The specific goals of the shared task are:

1. **Develop effective translation systems for low-resource Indian languages.** Models should handle small corpora, domain variation, and diverse scripts while producing faithful and stable translations.

2. **Evaluate cross-lingual transfer and parameter-efficient adaptation.** This includes approaches such as LoRA-based fine-tuning, multilingual joint training, retrieval augmentation, and hybrid SMT-NMT pipelines.

3. **Promote socially grounded and inclusive NLP.** Better translation systems can support information access in healthcare, education, disaster response, and public services, helping reduce the digital divide faced by tribal communities.

4. **Create open foundations for future research.** By releasing curated datasets and baseline systems, the shared task aims to enable long-term work in tribal language processing and cultural preservation.

2.1 Task Format

Participants were provided with four training files (one per language pair: Bhili–Hindi, Mundari–Hindi, Gondi–Hindi, and Santali–English), each containing parallel sentences with columns: **row_id** (unique identifier), a high-resource language column (`hindi` or `english`), and the corresponding tribal language column.

The test set provided unlabeled source sentences with columns: **row_id**, **source_sentence**, **source_lang**, and **target_lang**. Systems were expected to generate translations in both directions (high-to-low and low-to-high resource) and submit predictions with the format: **row_id**, **source_lang**, **source_sentence**, **target_lang**, and **target_sentence**.

2.2 Evaluation Metrics

Systems were ranked using a weighted composite score that balances word-level accuracy and character-level fluency. The final score is defined as:

Lang Pair	Columns
Hin ↔ Bhili	row_id, hindi, bhili
Hin ↔ Mundari	row_id, hindi, mundari
Hin ↔ Gondi	row_id, hindi, gondi
Eng ↔ Santali	row_id, english, santali
Test Data	row_id, source_sent

Table 1: Dataset Overview.

$$S = 0.6 \times \text{BLEU} + 0.4 \times \text{chrF}$$

BLEU (Papineni et al., 2002) receives higher weight (60%) due to its sensitivity to n-gram precision, while chrF (Popovic, 2015) contributes 40% to capture character-level morphology—crucial for languages like Mundari and Santali.

Scores were computed for each translation direction and aggregated with an additional directional weighting:

- **High-to-Low (60%):** Higher weight is assigned to translations *into* the tribal languages, reflecting the task's emphasis on accessibility for low-resource communities.

- **Low-to-High (40%):** Translations *from* the tribal languages are given slightly lower weight.

This composite metric ensures that systems are evaluated fairly across both lexical accuracy and morphological sensitivity.
Full Formula:

$$\text{Score} = 0.6\left[0.6 \sum_{d\in H\to L} \text{BLEU}_d + 0.4 \sum_{d\in L\to H} \text{BLEU}_d\right] + 0.4\left[0.6 \sum_{d\in H\to L} \text{chrF}_d + 0.4 \sum_{d\in L\to H} \text{chrF}_d\right].$$

Where $H \to L$ represents the high-to-low directions (e.g., Hindi → Bhili) and $L \to H$ represents the low-to-high directions.

3 Dataset

The shared task focuses on four low-resource tribal languages of India (see Table 1): Bhili, Mundari, Gondi, and Santali. All four have large speaker communities but extremely limited digital presence. Existing text is sparse, unaligned, or not machine readable, making these languages particularly challenging for machine translation. The dataset comprises 80,000 training pairs (20,000 per language pair) and 15,999 test sentences distributed across bidirectional translation tasks.

These language pairs represent a mix of typologically related and unrelated languages. Hindi, Bhili, and Mundari are written in Devanagari, though they belong to different language families (Indo-Aryan and Austroasiatic). Gondi, also written in Devanagari, is Dravidian. Santali, however, is an Austroasiatic language written in the Ol Chiki script and paired with English, making it one of the most linguistically distant settings in the task.

We provide a brief overview of the languages used in this challenge below:

Bhili (ISO 639-3: bhb) is a Western Indo-Aryan language spoken by approximately 13 million people across western and central India. Despite its large speaker base and proximity to major languages like Gujarati and Marathi, Bhili remains severely under-resourced in terms of digital content and NLP tools. The variety used in this dataset is the Jhabua dialect from Madhya Pradesh, written in Devanagari script.

Mundari (ISO 639-3: unr) is a North Munda language of the Austroasiatic family, spoken by approximately 1.6 million people primarily in Jharkhand and neighboring states. Although traditionally written in multiple scripts, we use Devanagari to maintain consistency with Hindi-aligned NLP pipelines. Mundari has complex word structures where multiple prefixes and suffixes combine to form single words, and suffers from limited parallel data and strong influence from regional dominant languages.

Gondi (ISO 639-3: gon) is a South-Central Dravidian language with approximately 3 million speakers distributed across central India. Due to language shift pressures, many speakers are shifting to regional dominant languages. We focus on the Devanagari-script variety to ensure compatibility with mainstream Hindi NLP tools and resources.

Santali (ISO 639-3: sat) is a major Munda language with over 7 million speakers across India, Bangladesh, Nepal, and Bhutan. Unlike the other three languages in this task, Santali is paired with English rather than Hindi. We focus on the Ol Chiki script, an indigenous writing system created in the 1920s that is increasingly used in education and modern publications. NLP resources for Ol Chiki remain extremely limited, particularly for parallel corpora.

3.1 Data Collection and Preprocessing

All parallel datasets were sourced from curated web content, Wikipedia articles, and resources provided by the Ministry of Tribal Affairs, Government of India. Each dataset follows a uniform structure with a unique row ID and a pair of parallel sentences. The data underwent a multi-stage preprocessing pipeline: sentence alignment, character normalization, deduplication of near-identical pairs, and filtering to retain only sentences between 6 and 80 words. Cosine similarity filtering removed pairs that were overly similar across source and target. For Bhili, expert-driven translation was employed to ensure accuracy and contextual quality. Tokenization was handled uniformly using SentencePiece (Kudo and Richardson, 2018).

For the English–Santali pair, the English source was produced by translating curated Hindi content using IndicTrans2 (Gala et al., 2023) to maintain thematic consistency across language pairs, with additional lowercasing applied to English text. The test set contains only the source sentence and language direction; participants must generate the target translation. It was created using stratified sampling across multiple domains to ensure coverage while avoiding repetition. All data is distributed under the Creative Commons BY-SA 4.0 license.

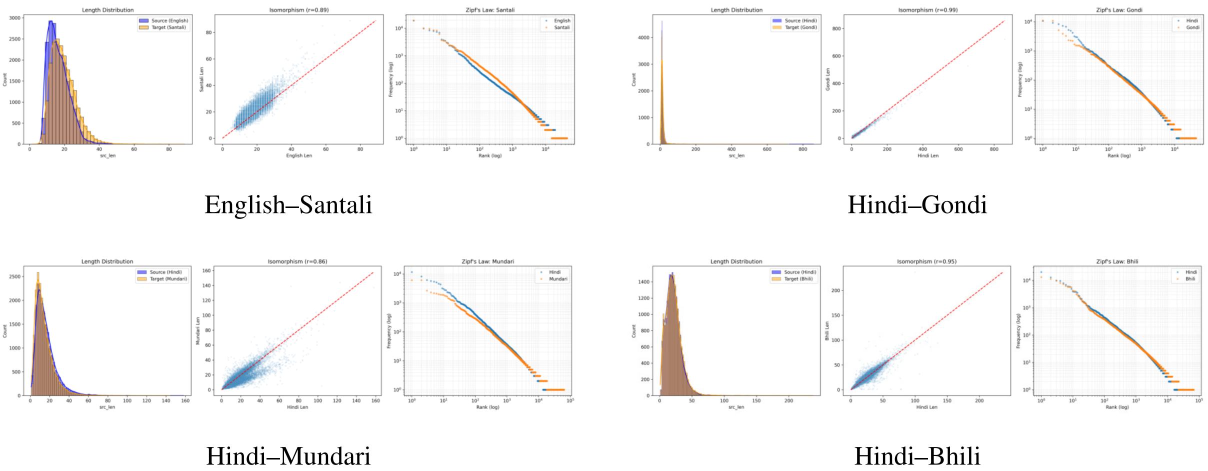
3.2 Dataset Statistics

Tables 2 and 3 summarize the corpus statistics. All training sets contain exactly 20,000 parallel pairs. Vocabulary analysis (Figure 1) reveals significant lexical sparsity, particularly in Bhili and Mundari, where target vocabularies exceed 60,000 tokens due to high morphological variation. Santali also exhibits high diversity driven by productive affixation in the Ol Chiki script. Structurally, Gondi remains the most compact (avg. 13.8 tokens), whereas Bhili presents the longest sequences in both source and target directions.

Overall, the dataset reflects real-world low-resource conditions: diverse scripts, high morphological richness, and large vocabularies relative to corpus size. These properties justify the need for specialized training strategies such as LoRA-based direction-specific fine-tuning, retrieval-augmented methods, and conservative decoding techniques.

3.3 Dataset Complexity Analysis

We present a comprehensive exploratory data analysis (EDA) of the results to understand the linguistic



English–Santali

Hindi–Gondi

Hindi–Mundari

Hindi–Bhili

Figure 1: EDA visualizations across the four language pairs.

Dataset	Src	Src Len	Src Voc	Tgt	Tgt Voc
santali-train	Eng	16.5	42114	San	44764
gondi-train	Hin	14.4	25133	Gon	44756
mundari-train	Hin	16.3	35373	Mun	63426
bhili-train	Hin	21.3	40636	Bhi	67019

Table 2: Training data statistics (20,000 pairs per language pair).

Source Lang	Count	Avg Len	Vocab Size
Bhili	1999	21.6	13125
English	2000	15.5	7810
Gondi	2000	6.9	5673
Hindi	6000	15.7	16860
Mundari	2000	14.1	11356
Santali	2000	14.4	7247

Table 3: Test set statistics

complexity of the corpus. Table 4 summarizes the metrics calculated using the NLLB tokenizer and standard lexical diversity measures.

Morphological Complexity

We computed *Token Fertility* - the average number of subword tokens required to represent a single word—as a proxy for morphological richness. As shown in Table 4, Santali and Mundari show extremely high fertility rates (3.02 and 2.55, respectively) compared to their source languages. This indicates highly agglutinative structures where a single word often translates to multiple tokens in the NLLB vocabulary, posing a significant challenge for the model.

Structural Divergence

We measured the Pearson correlation coefficient (r) between source and target sentence lengths.

- **High Isomorphism** ($r > 0.9$)**:** Hindi-Gondi ($r = 0.99$) and Hindi-Bhili ($r = 0.95$) show near-perfect length correlation. This suggests these languages share similar syntactic structures with Hindi, making them easier for models to align.
- **Divergence** ($r < 0.9$)**:** English-Santali ($r = 0.89$) and Hindi-Mundari ($r = 0.86$) show lower correlation, reflecting significant structural differences (e.g., SVO vs. SOV word order) that complicate translation.

Table 4: Linguistic Statistics of the Training Data. **TTR**: Type-Token Ratio (Lexical Richness). **Fertility**: Avg. subwords per word (NLLB tokenizer).

Pair	**Corr** (r)	**TTR**		**Fertility**	
		Src	Tgt	Src	Tgt
Eng-Santali	0.89	0.127	0.116	1.35	**3.02**
Hin-Gondi	**0.99**	0.088	0.162	1.39	2.32
Hin-Mundari	0.86	0.108	**0.223**	1.50	2.55
Hin-Bhili	0.95	0.095	0.155	1.43	1.73

4 Approaches and Results

The shared task attracted 18 teams who submitted solutions on Kaggle. Teams achieving a private leaderboard score above 150 were invited to submit challenge papers describing their approaches. Out of the seven eligible teams, four accepted this invitation and provided detailed system descriptions. Most teams started from multilingual pretrained models and fine-tuned them on the provided training data, adding specialized techniques to handle data scarcity, morphological complexity, and vocabulary mismatch. Table 5 presents the final rankings for the four teams that submitted papers.

Table 5: Complete final leaderboard ranked by private scores. Teams marked with * submitted challenge papers (invitation threshold: private score >150).

Rank	Team Name	Public	Private
1	SajayR*	319.39	**212.04**
2	HCMUS_PrompterX*	311.61	186.37
3	No	310.19	184.87
4	boy Magic	309.98	179.83
5	Shooting star*	216.04	179.49
6	VaibhavKanojia*	171.39	161.11
7	Shivansh Jha	193.36	153.90
8	Badr A	224.09	145.47
9	EROL	155.22	134.02
10	king	162.25	133.96
11	e0nia	143.89	131.56
12	c00k	290.30	111.47
13	Kabir Raj Singh	140.64	104.07
14	Daglox Kankwanda	122.63	96.22
15	Michael Ibrahim	92.00	87.17
16	Harsh Rajbhar	37.28	35.18
17	winner_can_exist	26.10	20.80
18	Code by Nadiia	4.22	6.28

4.1 System 1: LoRAs in All Directions

The winning team (Table 5, Rank 1) started from the widely used NLLB (Costa-jussà et al., 2022) backbone but later switched to the IndicTrans2 model (1.1B parameter version) (Gala et al., 2023). This decision was driven by a detailed analysis of tokenization behavior on the provided datasets. IndicTrans2 showed far superior fertility statistics, especially for Santali. For example, the model produced an average fertility of 1.44 tokens per character for Santali, compared to 3.07 for NLLB. Lower fertility indicates that the model represents the script more efficiently and reduces fragmentation, which is essential when dealing with agglutinative morphology. This gave IndicTrans2 a clear representational advantage.

4.1.1 Training Strategy

The team employed a three-stage "Saturate-then-Specialize" pipeline designed to balance multilingual generalization with task-specific specialization.

Tag-Only Preprocessing. To preserve contrastive linguistic properties often lost during script unification, the team utilized explicit language tags. Surrogate tags were mapped to the nearest vocabulary equivalent (e.g., Bhili to `mar_Deva`), enabling the model to internally align unseen languages with typologically similar ones without external transliteration.

Joint Fine-Tuning. The backbone was first fine-tuned on the union of all datasets to “saturate” the model with shared domain knowledge. This exposure to diverse lexical distributions stabilized the embedding space, improving robustness in data-sparse conditions.

Directional LoRA Adapters. To eliminate catastrophic forgetting, the backbone was frozen while training separate Low-Rank Adapters (LoRA) (Hu et al., 2022) for each direction ($r = 64, \alpha = 128$). This parameter isolation prevented cross-lingual interference (negative transfer) while allowing targeted adaptation for specific scripts and grammars.

4.1.2 Inference Strategy

To mitigate hallucinations, the system employed noisy channel reranking (Yee et al., 2019). Instead of relying solely on the forward probability $P(y|x)$, K candidates were generated and re-scored using a reverse model $P(x|y)$:

$$\hat{y} = \arg\max_{y \in \mathcal{Y}} [\alpha \cdot \log P(y|x) + \beta \cdot \log P(x|y)]$$

This objective acts as a semantic regularizer by rewarding translations from which the source sentence is reconstructible. Candidates that hallucinate or deviate in meaning yield low reverse probabilities and are effectively filtered out.

4.2 System 2: JHARNA MT

The runner-up team (Rank 2) identified hallucination as the primary failure mode in data-scarce settings and designed a hybrid architecture. By integrating retrieval-based memory and Statistical Machine Translation (SMT) with modern Neural Machine Translation (NMT), the system grounds neural generations in real training examples to ensure lexical fidelity.

4.2.1 Training Strategy

The pipeline prioritizes observed data via a two-tier retrieval module:

- **Exact Match:** Direct retrieval of training targets for inputs seen during training (8% of test data).
- **Fuzzy Match:** Retrieval of examples with edit distance ≤ 1 to ground common expressions.

For unseen inputs, candidates are generated by two distinct models:

- **SMT Branch:** An IBM Model 1 system (Brown et al., 1993) with diagonal alignment priors. This branch produces "literal" translations that are robust against NMT hallucinations.
- **NMT Branch:** An NLLB-200 (Costa-jussà et al., 2022) model fine-tuned with LoRA adapters (Hu et al., 2022) ($r = 16$) to provide fluency and context awareness.

4.2.2 Inference Strategy

To unify the branches, the system pools N hypotheses and applies Minimum Bayes-Risk (MBR) decoding (Kumar and Byrne, 2004). This selects the "consensus" translation that maximizes average similarity (BLEU/chrF) to all other candidates. This effectively filters out hallucinations, as they tend to diverge significantly from the stable SMT outputs.

The team observed complementary error modes: SMT is rigid but lexically safe, while NMT is fluent but prone to hallucination. The hybrid MBR approach successfully combines these strengths, yielding significant stability improvements over pure neural baselines.

4.3 System 3: Breaking Language Barriers

The Breaking Language Barriers team (Rank 5) adopted a data-centric strategy using NLLB-200 (Costa-jussà et al., 2022) as the primary model and mBART-50 (Liu et al., 2020) as a validator. Lacking native proficiency, the focus remained on detecting pathological model behaviors through statistical heuristics rather than linguistic intuition.

4.3.1 Training Strategy

To address data sparsity, all bidirectional pairs were concatenated into a single D_{unified} dataset. This doubled the effective training size and enforced a shared semantic space across all languages, encouraging cross-lingual alignment and reducing overfitting.

4.3.2 Inference Strategy

A “Safety Net” ensemble was implemented to filter catastrophic failures. If the primary model’s output length ratio was extreme (< 0.3 implying under-generation or > 3.0 implying hallucination), the system fell back to mBART predictions. This mechanism significantly improved robustness, boosting the private leaderboard score by over 5 points.

4.4 System 4: Divide and Translate

The Divide and Translate team (Rank 6) based their system on the hypothesis that multilingual models suffer from negative interference when dealing with languages that differ significantly in syntax. For example, English is an SVO language while Santali and the Hindi aligned languages are SOV. To avoid cross language interference, the team trained *separate decoders* for each direction while keeping a shared encoder.

4.4.1 Training Strategy

The shared encoder from NLLB-600M (Costa-jussà et al., 2022) was frozen completely. This prevented the degradation of the multilingual representations learned during large-scale pretraining. Only the decoder parameters were updated. This acted as a strong regularizer and reduced the risk of overfitting, which is common when training on small datasets.

4.4.2 Results

The approach produced very stable results with minimal public private score variance. However, the performance ceiling was lower than that of LoRA based or hybrid systems. The frozen encoder limited the model’s ability to adapt to the specific scripts and morphological patterns of the tribal languages.

5 Analysis and Key Findings

This section examines common patterns, challenges, and strategic decisions across the four submitted systems. We analyze the key factors that influenced performance, including choice of backbone models, approaches to mitigating hallucination, and how teams addressed script diversity and morphological complexity. These insights highlight both successful strategies and persistent challenges in low-resource machine translation.

5.1 Model Architecture Choices

A major distinction among the systems was the choice of backbone model. Three teams used NLLB (Costa-jussà et al., 2022), but the winning system built on IndicTrans2 (Gala et al., 2023). The differences stem from tokenization behavior. Many teams reported that NLLB struggled to tokenize Ol

Chiki and certain Devanagari variants. IndicTrans2, pretrained on a broader set of Indic languages, handled these scripts much more effectively.

Submissions also followed two main architectural strategies: **(1) Pure neural systems**, which rely solely on model scale, architecture, and decoding techniques; and **(2) Hybrid systems**, which combine SMT, retrieval, and neural models to address data scarcity. Hybrid approaches showed strong performance in low-resource settings, demonstrating that classical MT components remain useful when training data is limited.

5.2 Hallucination Mitigation Strategies

Hallucination was the dominant challenge of the task. Each system proposed a different mitigation strategy:

- **System 1:** Reverse model scoring (Noisy Channel).
- **System 2:** Consensus decoding (MBR).
- **System 3:** Length based fallbacks (Conservative Ensemble).
- **System 4:** Architectural constraints (Frozen Encoder).

The most effective strategies were the Noisy Channel and MBR approaches, both of which rely on generating multiple candidates and filtering them via external validation signals.

5.3 Script and Morphological Challenges

Santali, written in Ol Chiki, presented a major challenge for many backbones. Models without explicit support for the script required additional preprocessing or surrogate tags. IndicTrans2 had a clear advantage due to its broader coverage of Indic scripts.

Mundari has a very high type token ratio, more than double that of Hindi. This causes data sparsity and increases the difficulty of learning accurate word representations. System 2 used back translation to expand the dataset and expose the model to a wider range of morphological patterns.

5.4 Generalization and Overfitting

Some teams observed significant score drops between the public and private leaderboards. This indicates overfitting to specific patterns in the public test set. The Divide and Translate system showed the smallest gap, supporting their claim that encoder freezing works as a strong regularizer. The large gap between Public and Private scores (Table 5) highlights the challenge of generalization in low-resource settings.

6 Conclusion

The MMLoSo 2025 Shared Task is one of the first efforts to evaluate machine translation for India's tribal and severely low-resource languages. By releasing parallel datasets for Bhili, Gondi, Mundari, and Santali, the task provides a consistent benchmark for languages rarely covered in mainstream NLP. The task attracted 18 participating teams, with the top four achieving private scores above 150 and contributing detailed system descriptions. Teams employed low-resource-focused techniques such as retrieval augmentation, hybrid SMT–NMT systems, noisy-channel reranking, and direction-specific LoRA adapters to address challenges including script diversity, limited digital text, and rich morphology.

The results highlight the need for more documentation and larger datasets for tribal languages, which have large speaker communities but limited digital resources. By offering open data, standardized evaluation, and strong baselines, the shared task aims to support long-term research in inclusive and socially meaningful NLP. Future editions will expand toward multimodal translation, cross-dialect evaluation, and domain-specific benchmarks for governance, education, and cultural preservation.

7 Acknowledgements

We thank the Ministry of Tribal Affairs, Government of India, and especially Shri Vibhu Nayar for their guidance and support. We also acknowledge the Tribal Research Institutes of Odisha, Jharkhand, Chhattisgarh, and Madhya Pradesh, along with all translators and participants, for their essential contributions to this work.

References

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, Roman Ring, Eliza Rutherford, Serkan Cabi, Tengda Han, Zhitao Gong, Sina Samangooei, Marianne Monteiro, Jacob L. Menick, Sebastian Borgeaud, Andy Brock, Aida Nematzadeh, Sahand

Sharifzadeh, Mikolaj Binkowski, Ricardo Barreira, Oriol Vinyals, Andrew Zisserman, and Karén Simonyan. 2022. Flamingo: a visual language model for few-shot learning. In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*.

Peter F. Brown, Stephen Della Pietra, Vincent J. Della Pietra, and Robert L. Mercer. 1993. The mathematics of statistical machine translation: Parameter estimation. *Comput. Linguistics*, 19(2):263–311.

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, ACL 2020, Online, July 5-10, 2020*, pages 8440–8451. Association for Computational Linguistics.

Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Y. Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loïc Barrault, 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. *CoRR*, abs/2207.04672.

Jay P. Gala, Pranjal A. Chitale, Raghavan AK, 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. *Trans. Mach. Learn. Res.*, 2023.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. Lora: Low-rank adaptation of large language models. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net.

Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. The state and fate of linguistic diversity and inclusion in the NLP world. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 6282–6293. Association for Computational Linguistics.

Taku Kudo and John Richardson. 2018. Sentencepiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, EMNLP 2018: System Demonstrations, Brussels, Belgium, October 31 - November 4, 2018*, pages 66–71. Association for Computational Linguistics.

Shankar Kumar and William J. Byrne. 2004. Minimum bayes-risk decoding for statistical machine translation. In *Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics, HLT-NAACL 2004, Boston, Massachusetts, USA, May 2-7, 2004*, pages 169–176. The Association for Computational Linguistics.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven C. H. Hoi. 2023. BLIP-2: bootstrapping language-image pre-training with frozen image encoders and large language models. In *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, volume 202 of *Proceedings of Machine Learning Research*, pages 19730–19742. PMLR.

Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. 2020. Multilingual denoising pre-training for neural machine translation. *Trans. Assoc. Comput. Linguistics*, 8:726–742.

Wilhelmina Nekoto, Vukosi Marivate, Tshinondiwa Matsila, Timi E. Fasubaa, Taiwo Fagbohungbe, Solomon Oluwole Akinola, Shamsuddeen Hassan Muhammad, Salomon Kabongo Kabenamualu, Salomey Osei, Freshia Sackey, Rubungo Andre Niyongabo, Ricky Macharm, Perez Ogayo, Orevaoghene Ahia, Musie Meressa Berhe, Mofetoluwa Adeyemi, Masabata Mokgesi-Selinga, Lawrence Okegbemi, Laura Martinus, Kolawole Tajudeen, Kevin Degila, Kelechi Ogueji, Kathleen Siminyu, Julia Kreutzer, Jason Webster, Jamiil Toure Ali, Jade Z. Abbott, Iroro Orife, Ignatius Ezeani, Idris Abdulkabir Dangana, Herman Kamper, Hady Elsahar, Goodness Duru, Ghollah Kioko, Espoir Murhabazi, Elan Van Biljon, Daniel Whitenack, Christopher Onyefuluchi, Chris Chinenye Emezue, Bonaventure F. P. Dossou, Blessing K. Sibanda, Blessing Itoro Bassey, Ayodele Olabiyi, Arshath Ramkilowan, Alp Öktem, Adewale Akinfaderin, and Abdallah Bashir. 2020. Participatory research for low-resourced machine translation: A case study in african languages. In *Findings of the Association for Computational Linguistics: EMNLP 2020, Online Event, 16-20 November 2020*, volume EMNLP 2020 of *Findings of ACL*, pages 2144–2160. Association for Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, July 6-12, 2002, Philadelphia, PA, USA*, pages 311–318. ACL.

Maja Popovic. 2015. chrf: character n-gram f-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation, WMT@EMNLP 2015, 17-18 September 2015, Lisbon, Portugal*, pages 392–395. The Association for Computer Linguistics.

Ankita Shukla, Sandeep Kumar, Amrit Singh Bedi, and Tanmoy Chakraborty. 2025. Multimodal models for low-resource contexts and social impact 2025 machine translation challenge. Kaggle Competition. MMLoSo Workshop Co-located with IJCNLP-AACL 2025.

Lixu Sun, Nurmemet Yolwas, and Lina Jiang. 2023. A method improves speech recognition with contrastive learning in low-resource languages. *Applied Sciences*, 13(8):4836.

Kyra Yee, Yann N. Dauphin, and Michael Auli. 2019. Simple and effective noisy channel modeling for neural machine translation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, pages 5695–5700. Association for Computational Linguistics.