# TRANSLITERATION FOR INDIAN LANGUAGES

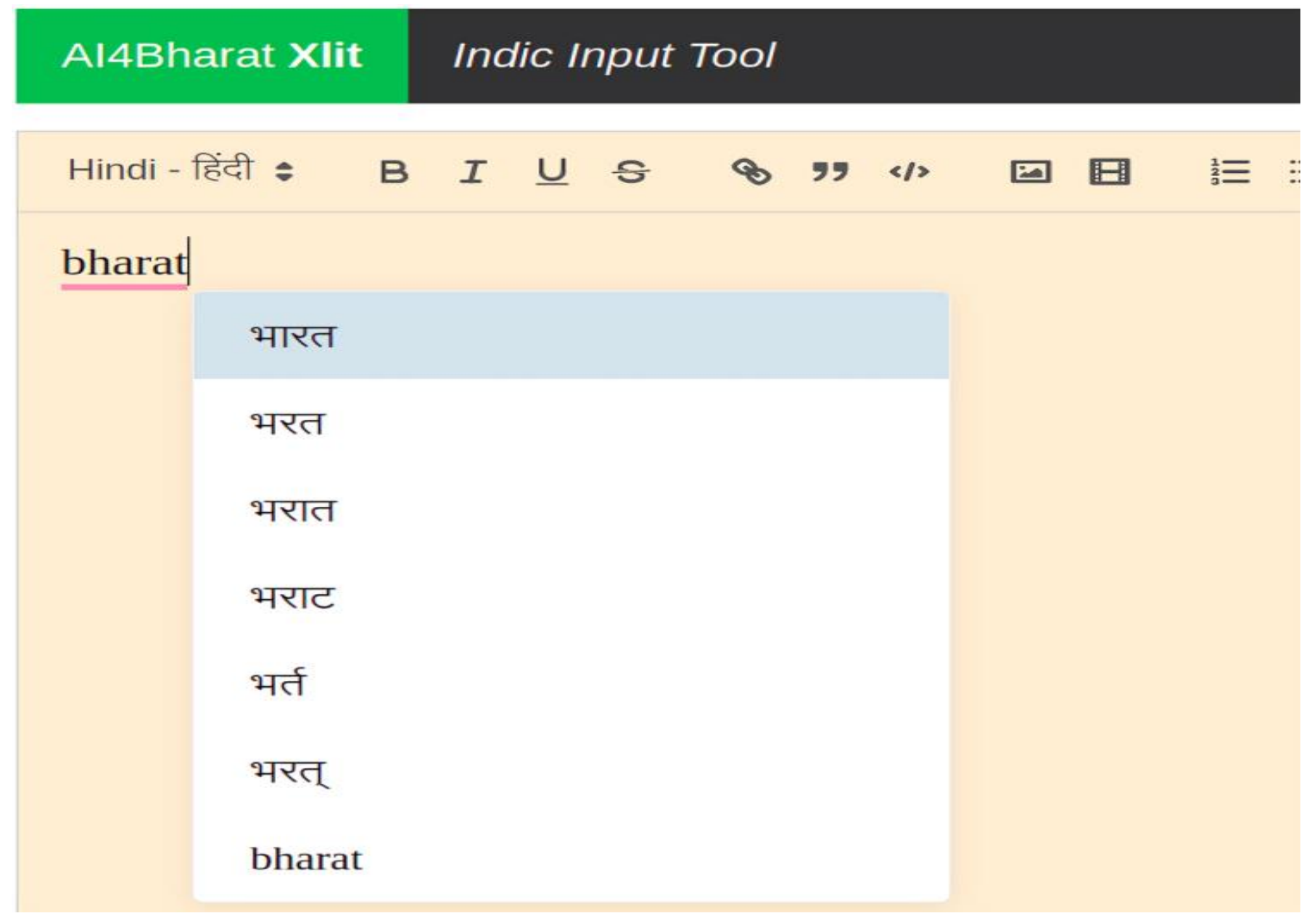
Yash Madhani, Sushane Parthan, Priyanka Bedekar, Gokul N C, Ruchi Khapra, Anoop Kunchukuttan, Pratyush Kumar, Mitesh M Khapra

## SUMMARY

- ***Aksharantar:*** **largest publicly available parallel transliteration corpus (26M word pairs)**
- ***IndicXlit:*** **Best open-source roman to native script transliteration model**
- **Diverse benchmark test set**
- **Dataset created by a combination of automated mining techniques and human generated transliterations.**

## What is Machine Transliteration?

- **Machine Transliteration refers to the automatic conversion of text in one script to text in another script *eg. Roman to Devanagari.***



bharat

भारत
भरत
भरात
भराट
भर्त
भरत्
bharat

https://xlit.ai4bharat.org/

## What is missing for Indian languages?

**Large scale transliteration training data**

**Diverse human annotated benchmark**

**Model across 22 Indian languages**

## Our Approach

**1** **Mine transliteration pairs for 21 Indian languages**

**2** **Collect manually annotated pairs**

**3** **Train a single Multilingual model**

## 1. Mining Transliteration Pairs

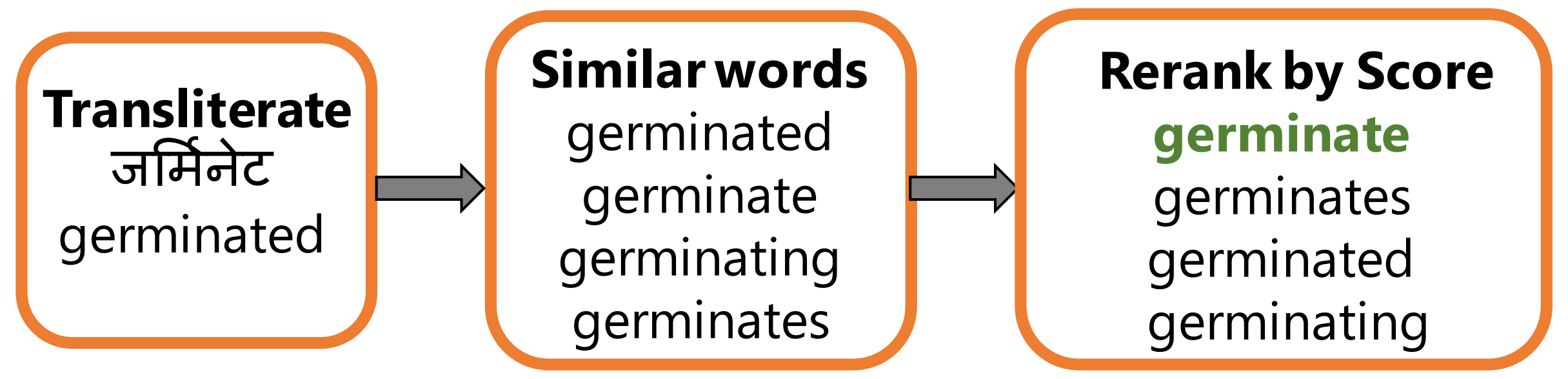
**26 M** **transliteration pairs mined**

**21** **Indian languages**

- **Mining from Samanantar**

**India** will wear the orange **jersey** in **match** against **England** on **June** 30 ⟷ टीम **इंडिया** 30 **जून** को विश्व कप **मैच** में **इंग्लैंड** के खिलाफ नारंगी **जर्सी** में खेलेगी

- **Mining from IndicCorp**



*Constitutes 90% of total mined data*

## 2. Large-Scale Manual Data Collection

**Collected 103k pairs in 19 languages**

**2-step maker-checker process**

**Different types of Indic words Romanized**

**68 language experts across India**

**Built-in Automatic Transliteration validator**

*Coverage of low resource langs. and diverse benchmark*

## 3. Train Multilingual Transliteration Model

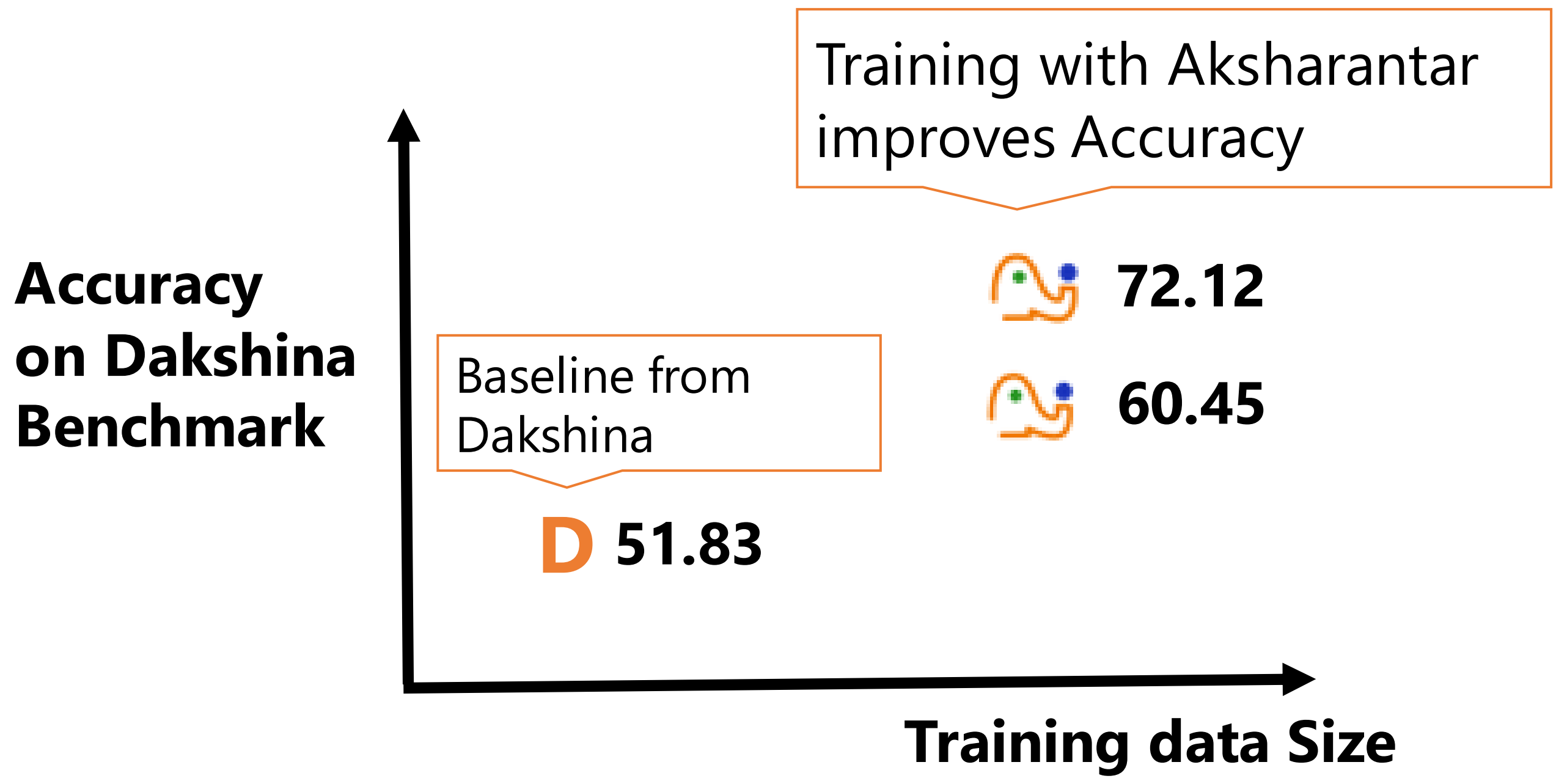
**A Multilingual model trained on 26M word pairs**

**Reranking using frequency in IndicCorp**

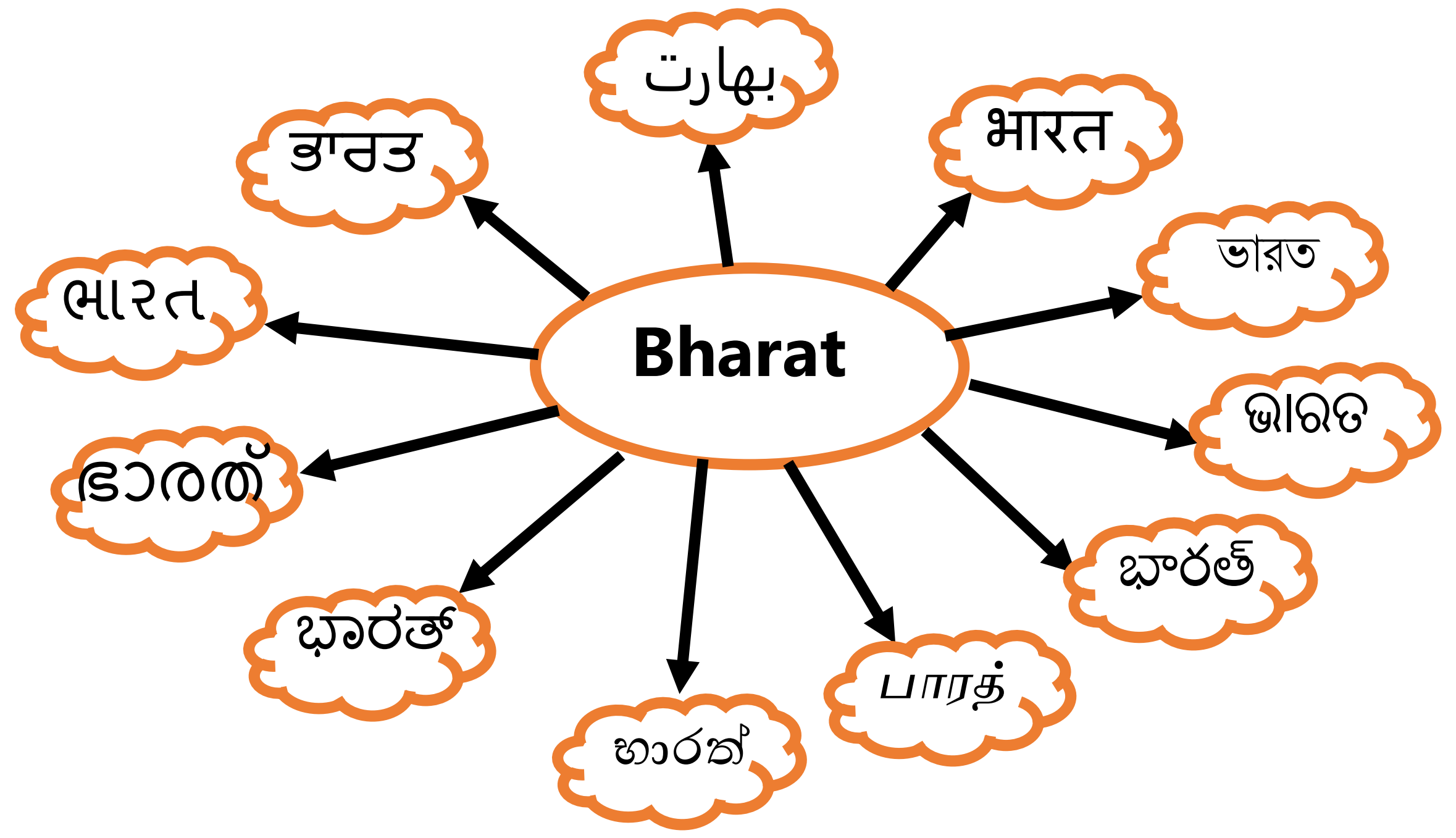
**Covers 21 Indian languages**

## Results

- **Model achieves state-of-the-art performance on Dakshina benchmark**



## Model Outputs



### OUR PLAN AHEAD

**Extend model support to include more Indian languages and dialects**

**Improve model efficiency using non-autoregressive (NAR) generation**

**Integrate transliteration model with swipe-based keyboard**

## ACKNOWLEDGEMENTS

We would like to thank the Nilekani Philanthropies for their generous grant which helped in setting up the "Nilekani Centre at AI4Bharat, IIT Madras" to support our students and research staff, as well as data and computational requirements. We would like to thank The Ministry of Electronics and Information Technology for its grant to support the creation of datasets and models for Indian languages under its ambitious Digital India Bhashini project. We would also like to thank the Centre for Development of Advanced Computing, India (C-DAC) for providing access to the Param Siddhi supercomputer for training our models. Lastly, we would like to thank Microsoft for its grant to create datasets and tools for Indian languages.

**The focus of AI4Bharat, an initiative of IIT Madras, is on building open-source language AI for Indian languages, including datasets, models, and applications.**

https://ai4bharat.iitm.ac.in/transliteration
https://github.com/AI4Bharat/IndicXlit
Contact: Yash Madhani, Sushane Parthan