

AI4Bharat: Transforming the Landscape of Indian Language Technology

Anoop Kunchukuttan

AI4Bharat



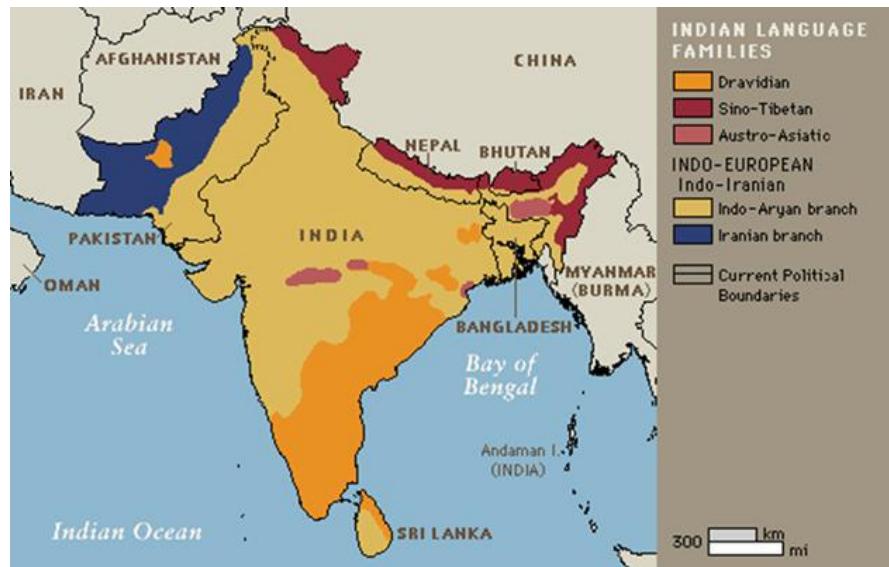
<https://ai4bharat.iitm.ac.in>

Mission Statement

Bring parity with English
in AI tech for Indian languages
with open data and open source contributions

*Build an ecosystem of datasets, models, partners
and stakeholders to advance IndicNLP*

Usage and Diversity of Indian Languages



- 4 major language families
- 22 scheduled languages
- 125 million English speakers
- 8 languages in the world's top 20 languages
- 30 languages with more than 1 million speakers

Sources: Wikipedia, Census of India 2011



Internet User Base in India (in million)
Source: Indian Languages:
Defining India's Internet KPMG-Google Report 2017

We are faced with a huge data skew

**Raw Text
Corpora**

*Wikipedia
articles*



**6m
150k**

Parallel Corpora

*Sentence
pairs*



**500m
1.5m**

NER Corpora

Tokens



**200k
40k**

QA

*Question-Answer
Pairs*



**100k
4.6k**

*How do build language technology solutions for Indian languages
that are of high-quality &
serve the use-cases of interest to us?*

The Pillars of Indian Language Technology Development

Collaboration
and Openness

Data Curation

Multilinguality
& Relatedness

Efficiency

Evaluation

Cultural
Awareness

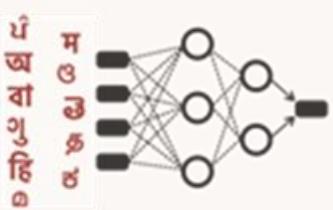
Systems &
Solutions

Tooling &
Infrastructure

The “Recipe” for Language Scalability



Crawl
monolingual
corpora



Pretrain a
multilingual
model



Mine Labelled
datasets



Fine-tune using
labelled data



Create benchmarks
for evaluation

Building Large-scale Indic Datasets

Raw Text Data is a critical resource

Why do we need raw text?

Compiles the collective knowledge of the web!

- Modern LLMs are trained on 10s of trillions of tokens
- Most of the data is in English

Captures language-specific Cultural Knowledge

A feeder resource for extracting many other resources

LM Training Corpora

Parallel Translation Corpora
Parallel Transliteration Corpora
Text Classification
NER Corpora
Language Generation

Challenges in building high-quality corpora

- *Large-scale crawling and processing*
- *Source identification*
- *Language identification*
- *Low-quality pages like MT*
- *Page content extraction*
- *Content Moderation*

IndicCorp v1

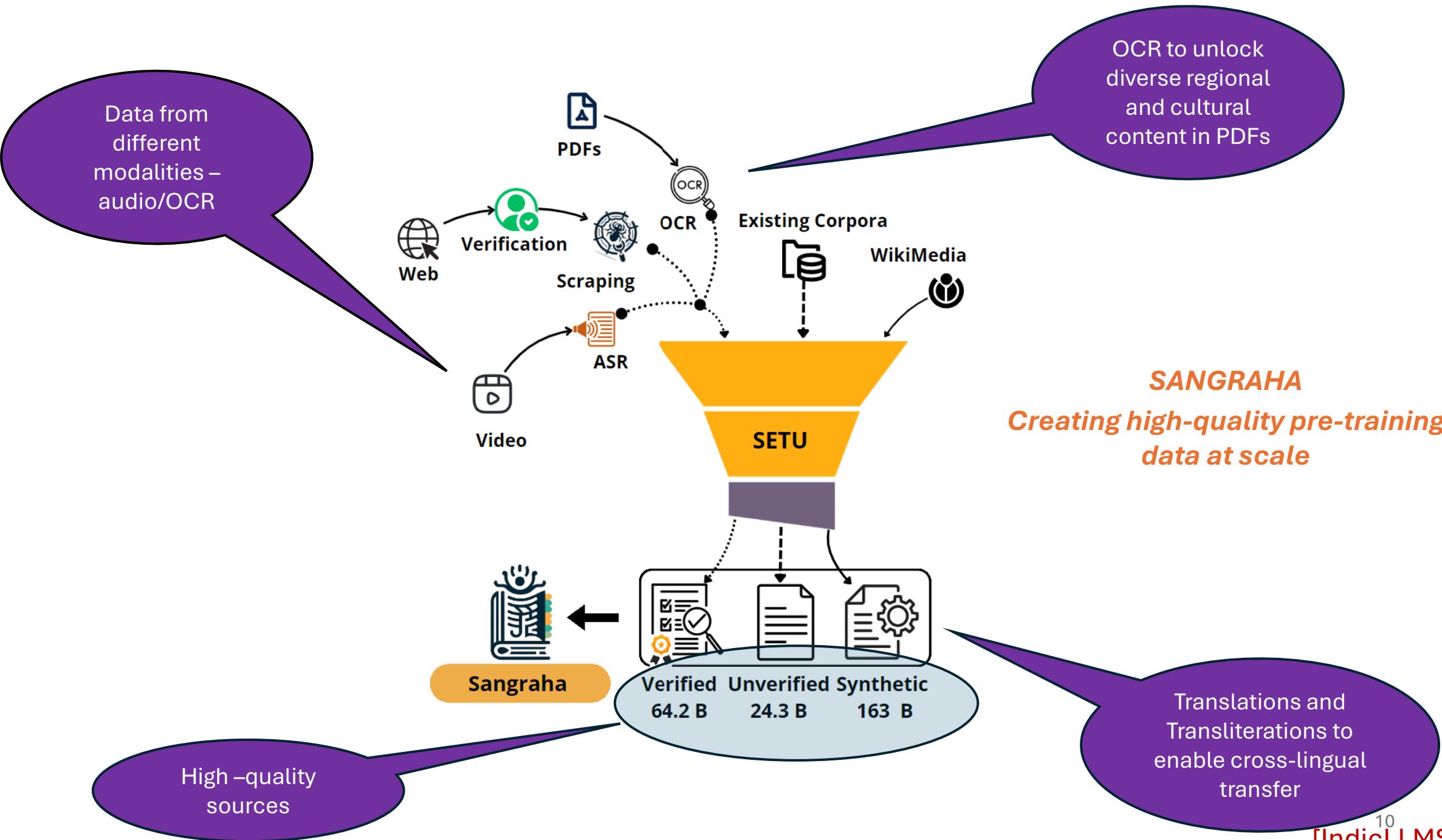
*Sentence-level
Web-sources*

IndicCorp v2

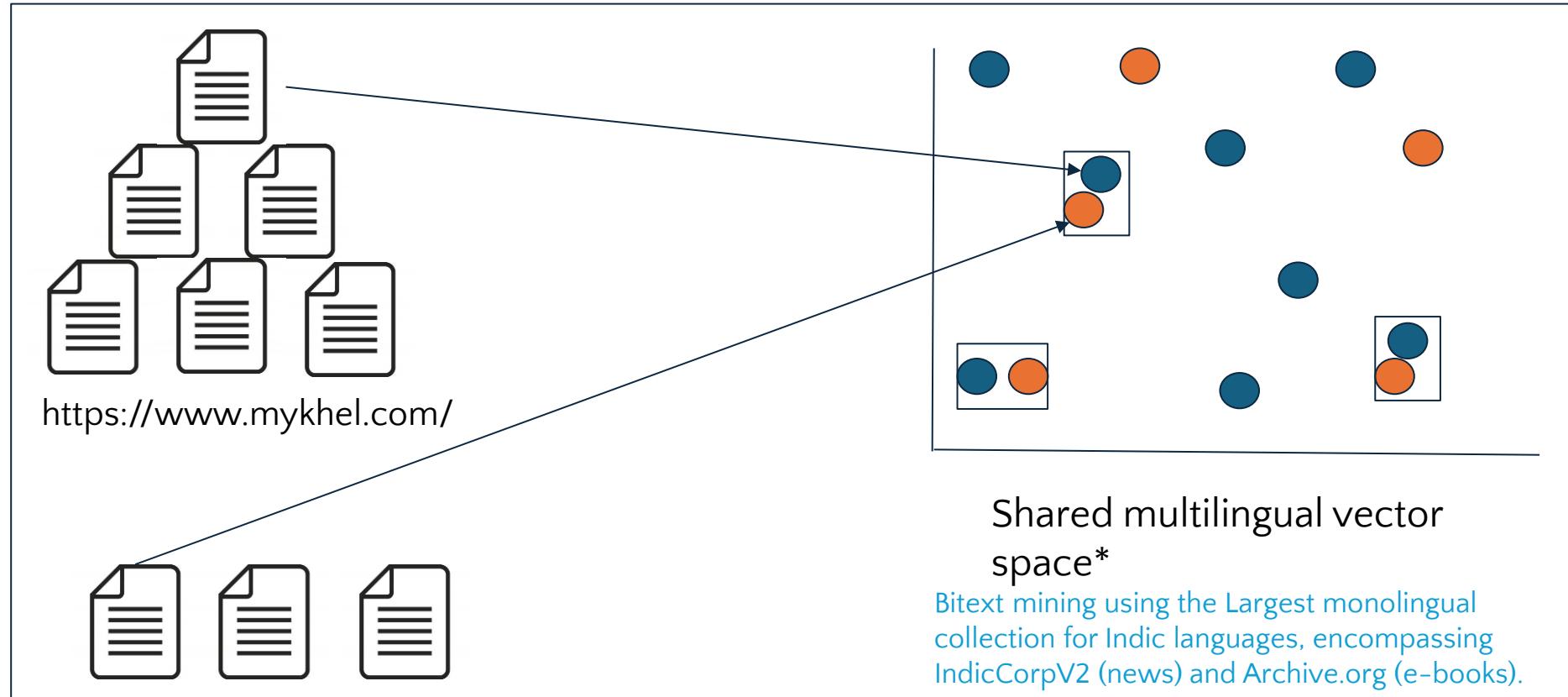
*Larger corpora
Larger language coverage*

Sangraha

*Document level
Diverse sources
Better filtering*



Mining Data for training Translation Systems



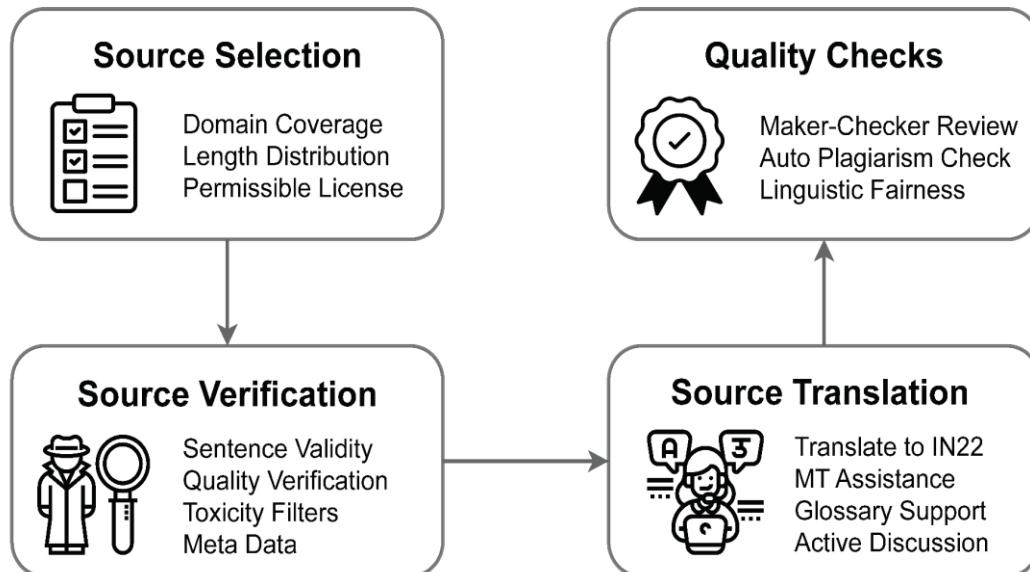
BPCC Mined Corpus: Mined 232 M sentence pairs across 12 Indian languages

*Fangxiao Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. Language-agnostic BERT sentence embedding, ACL 2022.

Expert Annotation: Translation

Boost model quality with high-quality expert annotations!

- High Quality translations can boost translation quality on fine-tuning
- Only source for very low-resource languages



Shoonya Organization Projects Datasets Analytics Admin

Back to Project

Notes Glossary

Auto-save enabled for this scenario.

	Draft	Next
#2054854	Ishvinder Sethi #883865 1/1	◀ ▶ × ↻ ↻ Skip Update

Source sentence
The Nilamata Purana is believed to have been commissioned by Durlabhavardhana.

Assamese translation
বিশ্বাস কৰা হয় যে নীলামাতা পুরাণটো দুর্লভবৰ্ধনৰ দ্বাৰা আৰম্ভ হৈছিল।

Machine translation
বিশ্বাস কৰা হয় যে নীলামাতা পুরাণটো দুর্লভবৰ্ধনৰ দ্বাৰা আৰম্ভ হৈছিল।

Context
The Nilamata Purana is believed to have been commissioned by Durlabhavardhana. [6][11] The Vishnudharmottara Purana, was crafted around the same times. [8][11] A famed patron of arts, Lalitaditya invited scholars from abroad to his court and promoted study of religions. [2]

Task #2054854

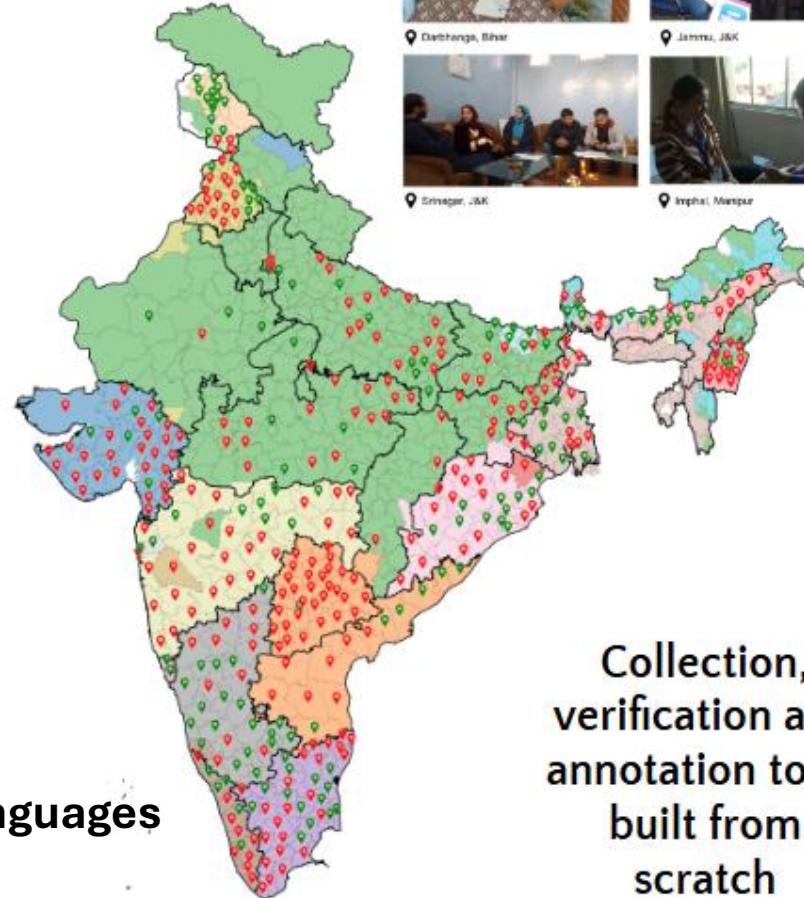
- Need processes in place to ensure high quality
- Provide tools to make translators productive

BPCC-Human Corpus: Created 800 K high-quality sentence pairs across 22 Indian languages

Expert Annotation: Automatic Speech Recognition

IndicVoices Corpus

22 languages



Collection,
verification and
annotation tools
built from
scratch



Capturing the voice of India

400+ Districts 1893 Personnels

12000
HOURS
COLLECTED



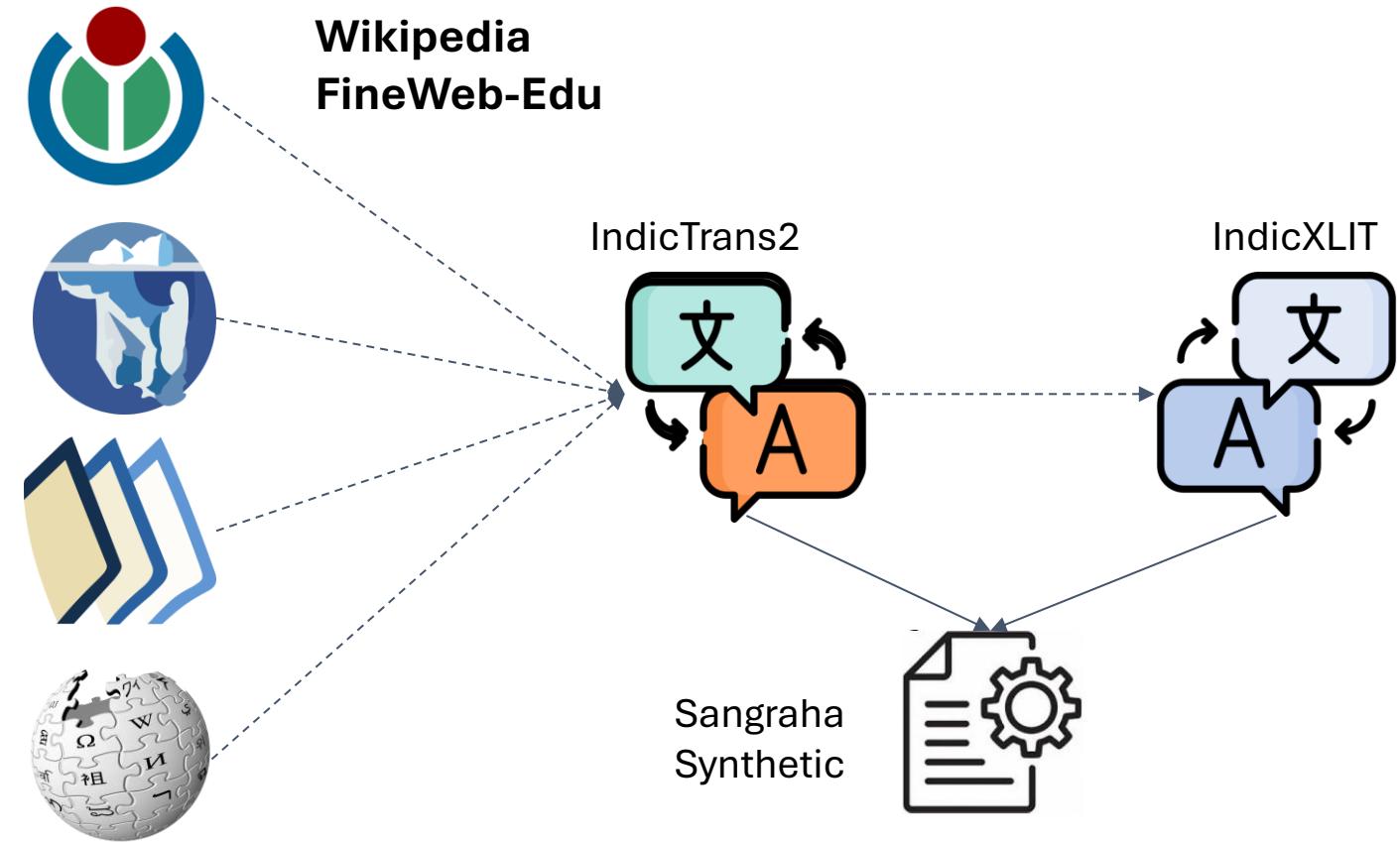
Generate Multilingual Data

*High-quality, efficient,
contextual, format-preserving
translation pipelines needed*

Huge disparity in digital knowledge between English and any other language

Quick Alternatives →

- **Translate** knowledge rich corpora to infuse knowledge in non-English languages
- **Transliterate** corpora to encourage cross-lingual transfer with English

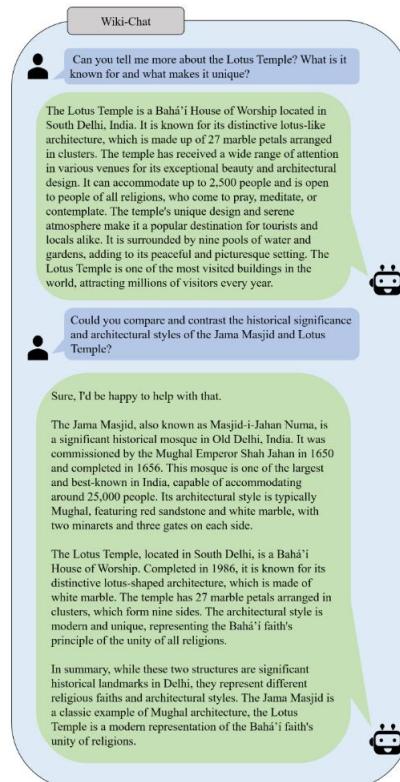
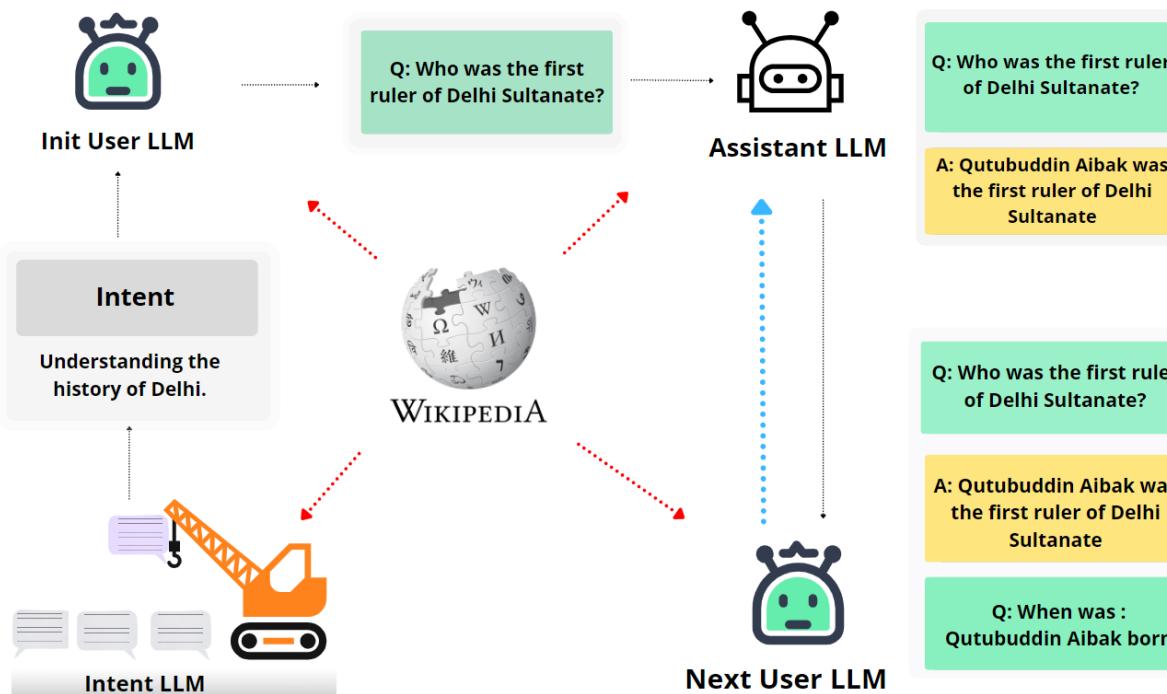


Creativity is the limit

Synthesize Culturally Relevant IFT Data

Translate the conversations from English

Use multiple English LLMs along with Wikipedia context to simulate conversations on topics of interest



(a) English

(b) Telugu

A Glance at AI4Bharat Technology

Publications: <https://ai4bharat.iitm.ac.in/publications>

Github: <https://github.com/AI4Bharat>

Huggingface: <http://huggingface.co/AI4Bharat>

Building the **foundation** to enable machines understand the rich and diverse lives of Indians.

IN22

ASR

OCR

MT

TTS

XLIT



DEPLOYMENTS



ROBUST
EVALUATIONS



MULTILINGUAL
MODELS

FOUNDATIONAL
DATASETS

Impact enabling
Inclusion

ALL
assisted language learning



NLP Infrastructure: Raw corpora



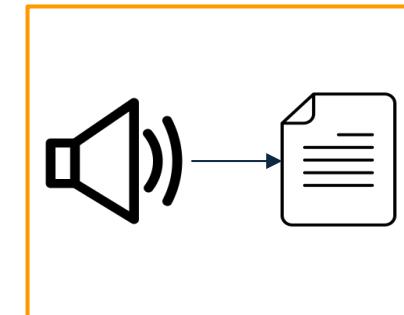
IndicCorp

Large Monolingual corpora
20B tokens, 22 languages



Sangraha

Large Document
Monolingual corpora
100B tokens, 22 languages



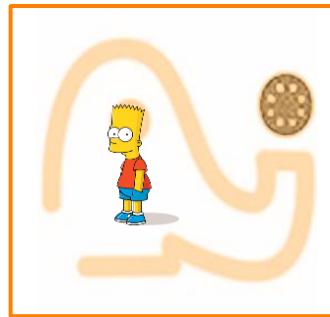
MahaDhwani

Raw speech corpora
(279k hours, 22
languages)

NLP Infrastructure: Language models

IndicFT

(word embeddings)



IndicWav2Vec

(Pre-trained speech model)

IndicBERT

(encoder LM)

IndicBART

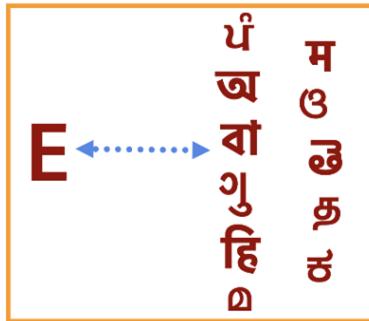
(seq2seq LM)

Airavata

(Finetuned LLM)

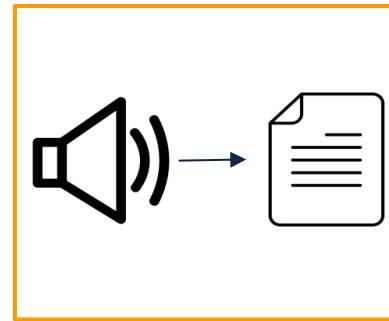
Compact pre-trained models for NLU & NLG

Data for various foundational tasks



BPCC

Parallel corpus,
translation models
between English & 22
Indic languages



Shrutilipi, IndicVoices &

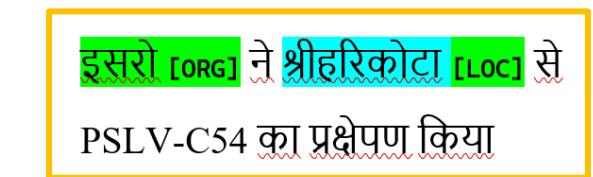
KathBath

ASR datasets for 22
Indian languages



Aksharantar

Transliteration datasets
for 20 Indic languages



Naamapadam

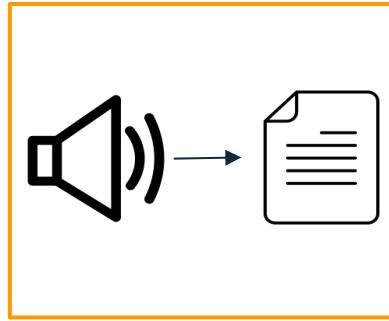
Datasets and models for
Named Entity
Recognition in 11 Indian
languages

Models for foundational tasks



IndicTransv2

Parallel corpus,
translation models
between English & 22
Indic languages



IndicConformer

ASR models for 22 Indian
languages



IndicXlit

Transliteration Models
for 20 Indic languages



IndicNER

Models for Named Entity
Recognition in 11 Indian
languages

Tools and Infrastructure

Shoonya

Shoonya is an open-source platform to improve the efficiency of language work in Indian languages with AI tools and custom-built UI interfaces and features. This is a key...

[Learn more →](#)

Chitralekha

Chitralekha is an open-source video transcreation platform for video subtitling, translation and voice-over generation across various Indic languages, using ML models suppo...

[Learn more →](#)

Kathbath

Kathbath is an open-source crowdsourcing toolkit designed for data collection in low-resource communities. It operates offline and syncs with a backend server when..

[Learn more →](#)

Anuvaad

Anuvaad is an open source judicial domain, document-translation platform to translate judicial documents at scale. Separate instances of Anuvaad are deployed..

[Learn more →](#)

Anudesh

Anudesh is an open-source platform dedicated to advancing the development of state-of-the-art Large Language Models for Indian languages.

[Learn more →](#)

Indic Glossary Explorer

Indic Glossary Explorer is an open source service to store and explore relevant Indic glossary which are domain specific. The service also provides the capabilities for glossa...

[Learn more →](#)

Standard Evaluation Benchmarks



IndicGLUE

In-language Benchmarks for Natural Language Understanding

IndicXTREME

Cross-lingual Benchmarks for Natural Language Understanding



Indic NLG Suite

Benchmarks for Natural Language Generation

Datasets for tasks like headline generation, paraphrase generation, etc



Indic SUPERB

Benchmarks for Speech Language Understanding

Vistaar

Datasets for tasks like ASR, speaker verification, speaker identification, LID etc



MILU

Benchmarking LLMs

IndicBIAS

Datasets for LLM evaluations on various aspects



Publications: <https://ai4bharat.iitm.ac.in/publications>



Github: <https://github.com/AI4Bharat>



Huggingface: <http://huggingface.co/AI4Bharat>

You can try out models at

<https://ai4bharat.iitm.ac.in>

IndicTrans2

Conference : TMLR Downloads last month : 20859

IndicTrans2 is the first open-source transformer-based multilingual NMT model that supports high-quality translations across all the 22 scheduled Indic languages — including multiple scripts for low-resource languages like Kashmiri, Manipuri and Sindhi. It adopts script unification wherever feasible to leverage transfer learning by lexical sharing between languages. Overall, the model supports five scripts Perso-Arabic (Kashmiri, Sindhi, Urdu), Oi Chiki (Santali), Meitei (Manipuri), Latin (English), and Devanagari (used for all the remaining languages).

Github Paper Colab Huggingface

Select Service: *

ai4bharat/indctrans--gpu-t4

Select Source Language: *

Kashmiri

Select Target Language: *

Kashmiri

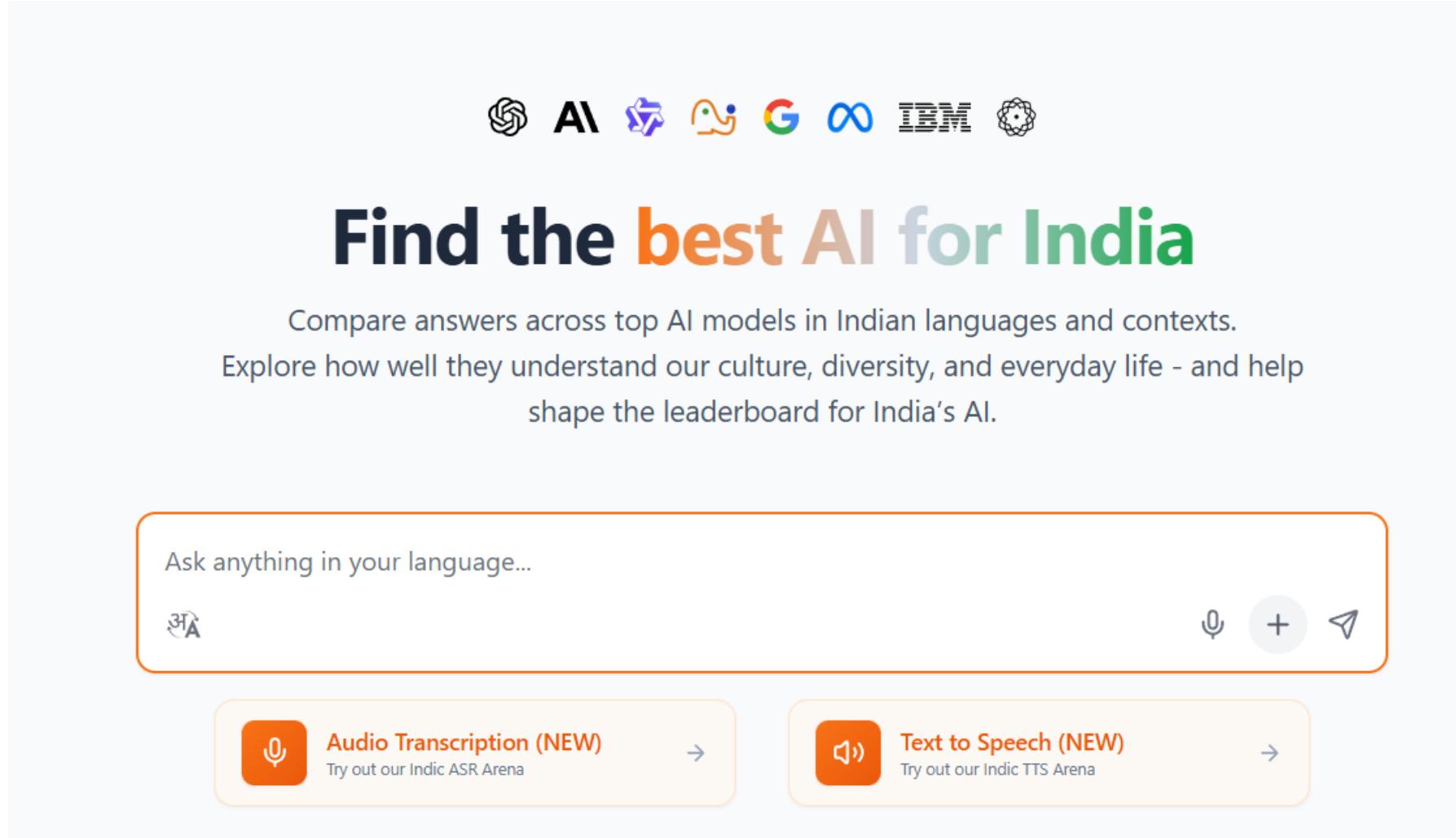
Enable Transliteration: *

Allow the AI to be improved by usage analysis. *

Translate

Indic LLM Arena

<https://arena.ai4bharat.org>



The screenshot shows the homepage of the Indic LLM Arena. At the top, there are logos for various AI models: Qwen, AI21, Qwen+, Qwen, Google, Qwen, IBM, and Qwen. Below these, the main heading reads "Find the **best AI for India**". Subtext explains: "Compare answers across top AI models in Indian languages and contexts. Explore how well they understand our culture, diversity, and everyday life - and help shape the leaderboard for India's AI." A large input field is centered, with the placeholder "Ask anything in your language..." and a microphone icon. To the right of the input field are three icons: a microphone, a plus sign, and a right-pointing arrow. At the bottom, there are two buttons: "Audio Transcription (NEW)" with a microphone icon and "Text to Speech (NEW)" with a speaker icon. Both buttons have the text "Try out our Indic [TTS/ASR] Arena" below them, followed by a right-pointing arrow.

Find the **best AI for India**

Compare answers across top AI models in Indian languages and contexts.

Explore how well they understand our culture, diversity, and everyday life - and help shape the leaderboard for India's AI.

Ask anything in your language...

AI21 Qwen Qwen+ Qwen Google Qwen IBM Qwen

Audio Transcription (NEW)

Text to Speech (NEW)

AI for Education in India

- Centre of Excellence (CoE) in Artificial Intelligence for Education at IIT Madras
- Develop India-first, population-scale AI based education solutions
- We are looking for people in different areas of engineering and Applied AI: [[Details Here](#)]

Thanks

anoop.kunchukuttan@gmail.com

<http://huggingface.co/AI4Bharat>

<http://anoopkunchukuttan.github.io>

Acknowledgments: All my collaborators, colleagues and students at AI4Bharat