

AI4Bharat: Transforming the Landscape of Indian Language Technology

Anoop Kunchukuttan

AI4Bharat



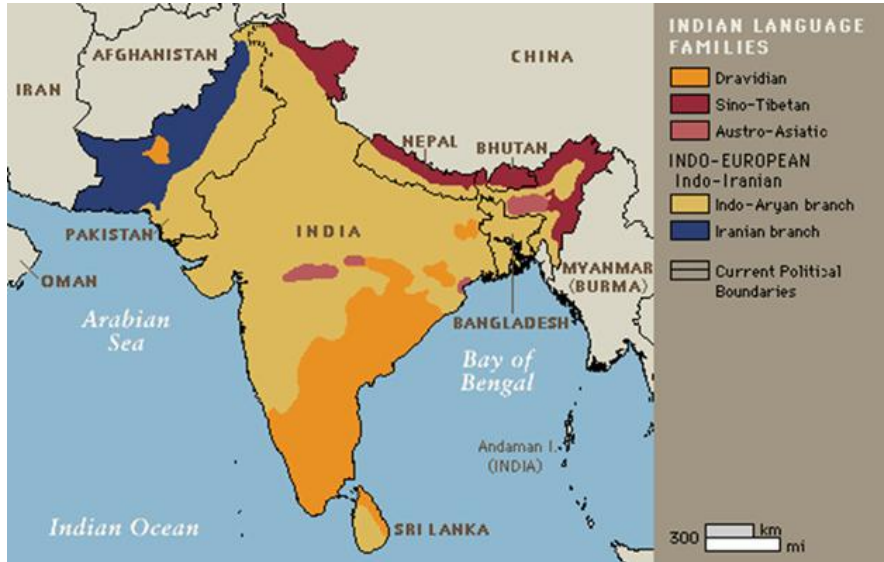
<https://ai4bharat.iitm.ac.in>

Mission Statement

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

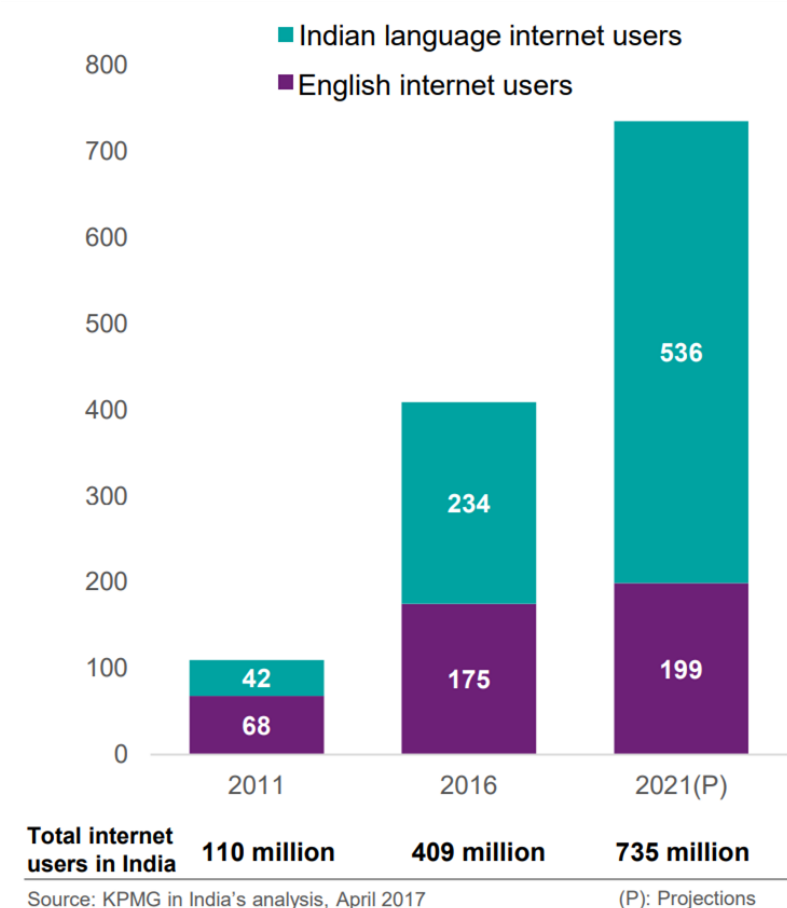
*Build an ecosystem of datasets, models, partners
and stakeholders to advance IndicNLP*

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

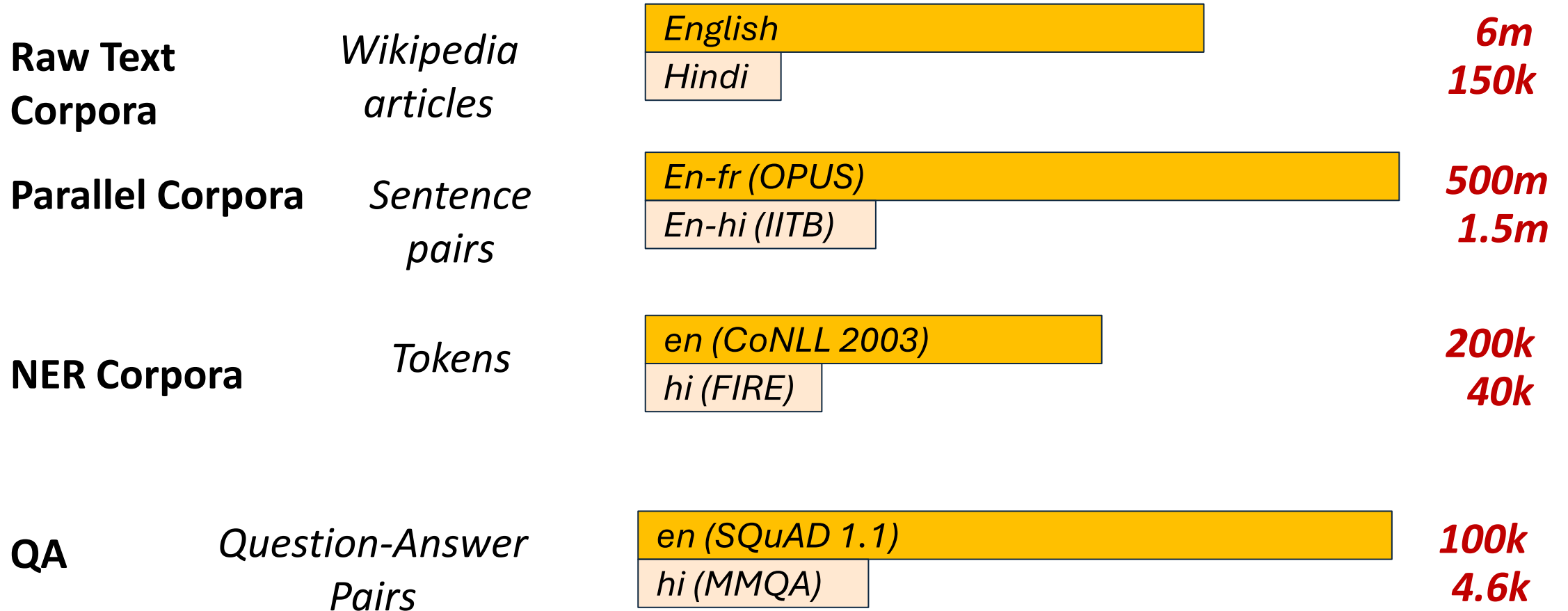
Sources: Wikipedia, Census of India 2011



Internet User Base in India (in million)

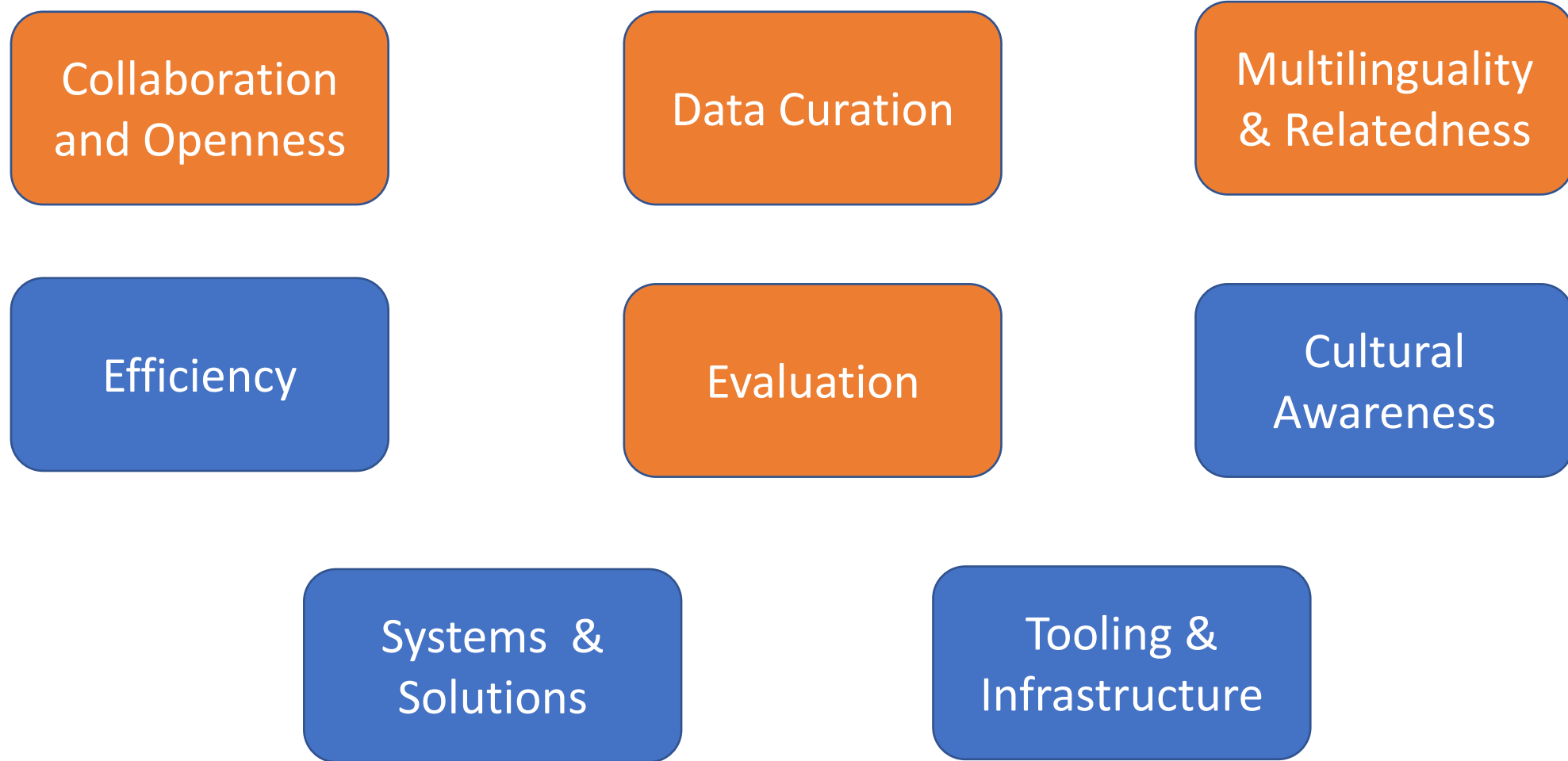
Source: Indian Languages:
Defining India's Internet KPMG-Google Report 2017

We are faced with a huge data skew



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

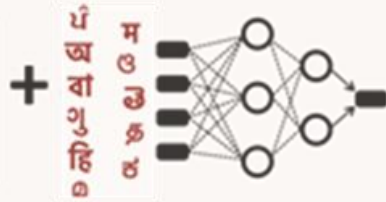
The Pillars of Indian Language Technology Development



The “Recipe” for Language Scalability



Crawl
monolingual
corpora



Pretrain a
multilingual
model

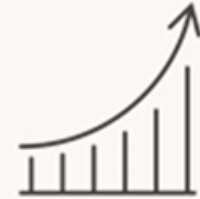


Mine Labelled
datasets



+

Fine-tune using
labeled 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

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*

LM Training Corpora

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

IndicCorp v1

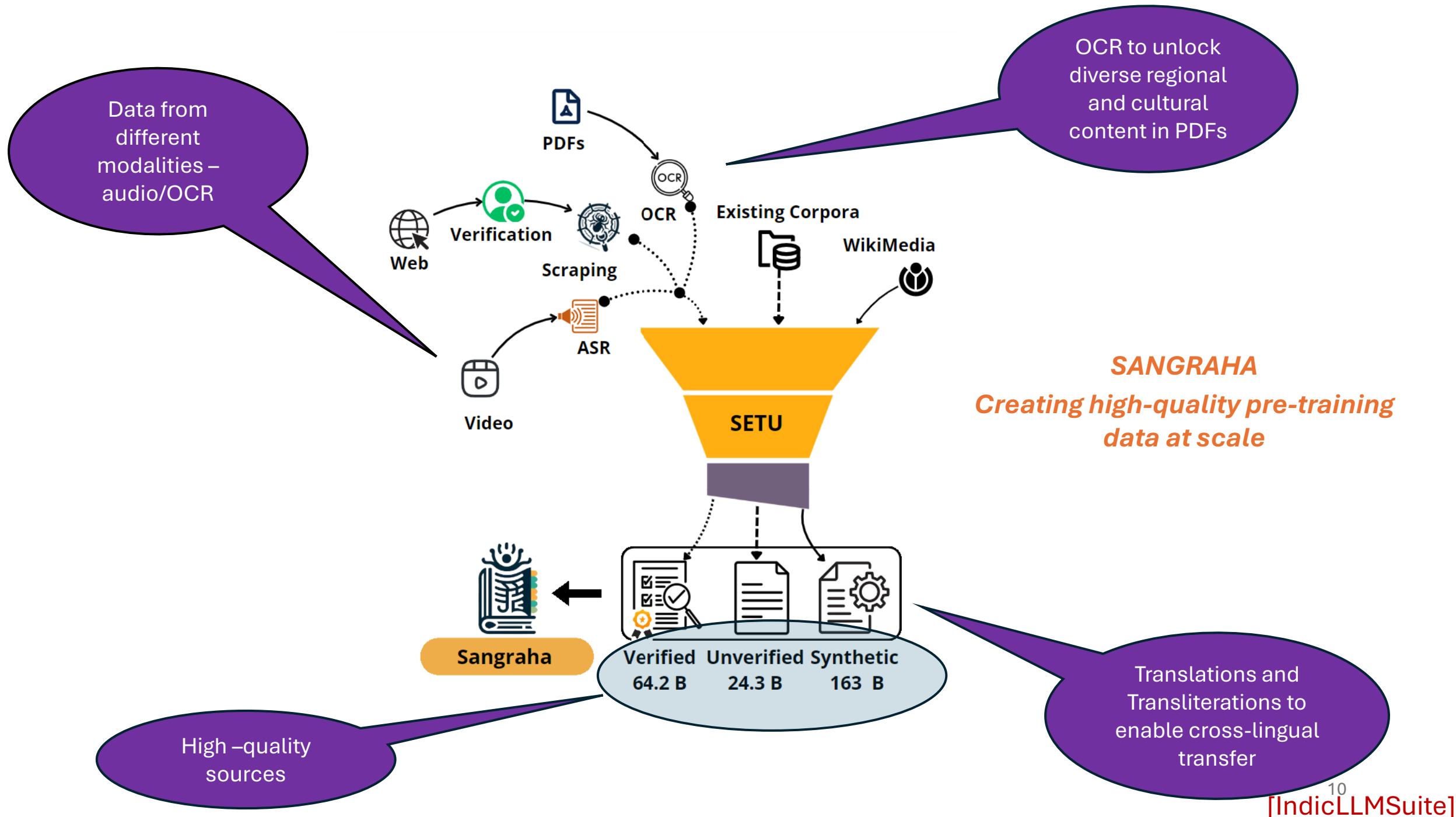
*Sentence-level
Web-sources*

IndicCorp v2

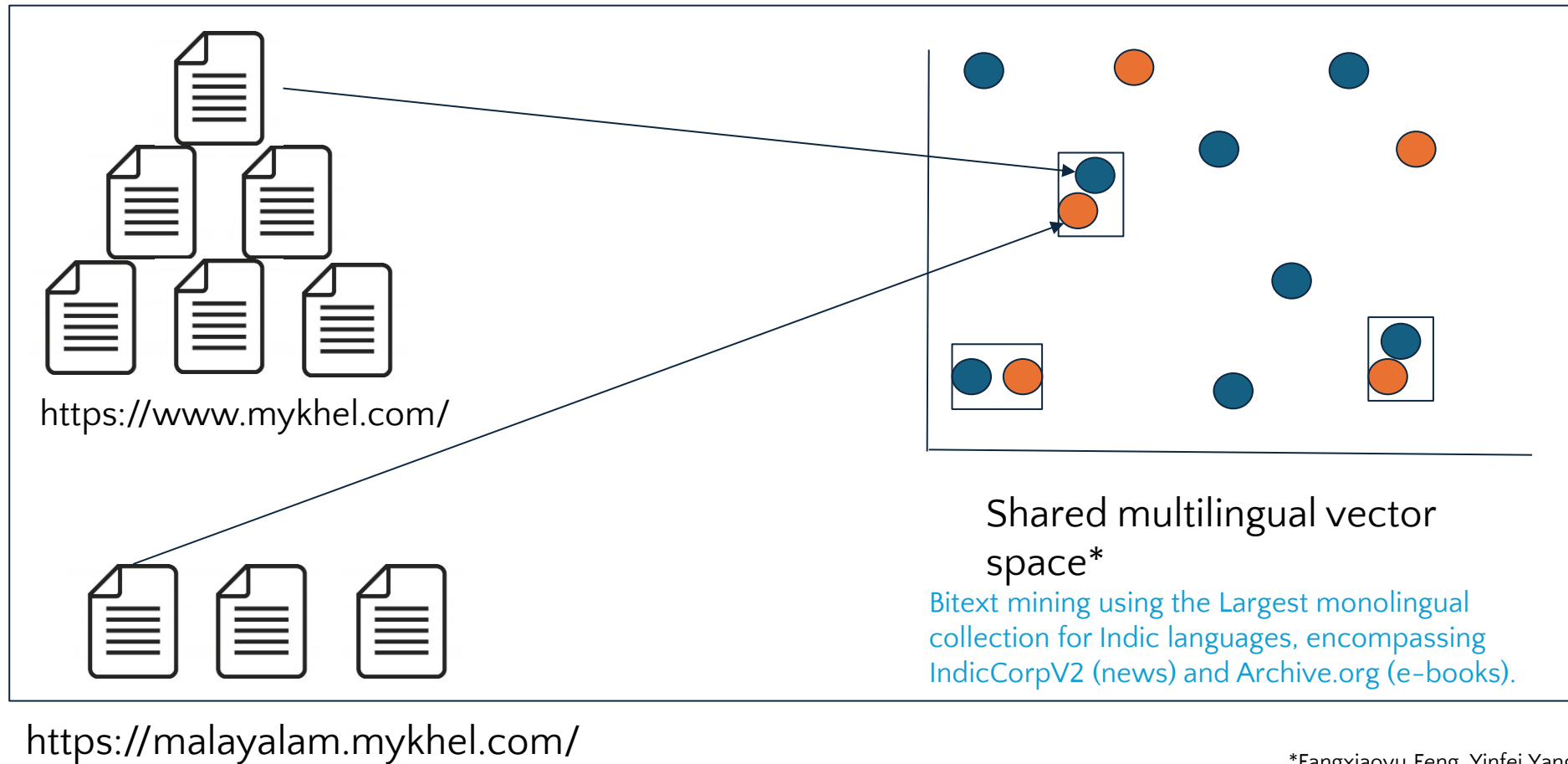
*Larger corpora
Larger language coverage*

Sangraha

*Document level
Diverse sources
Better filtering*



Mining Data for training Translation Systems



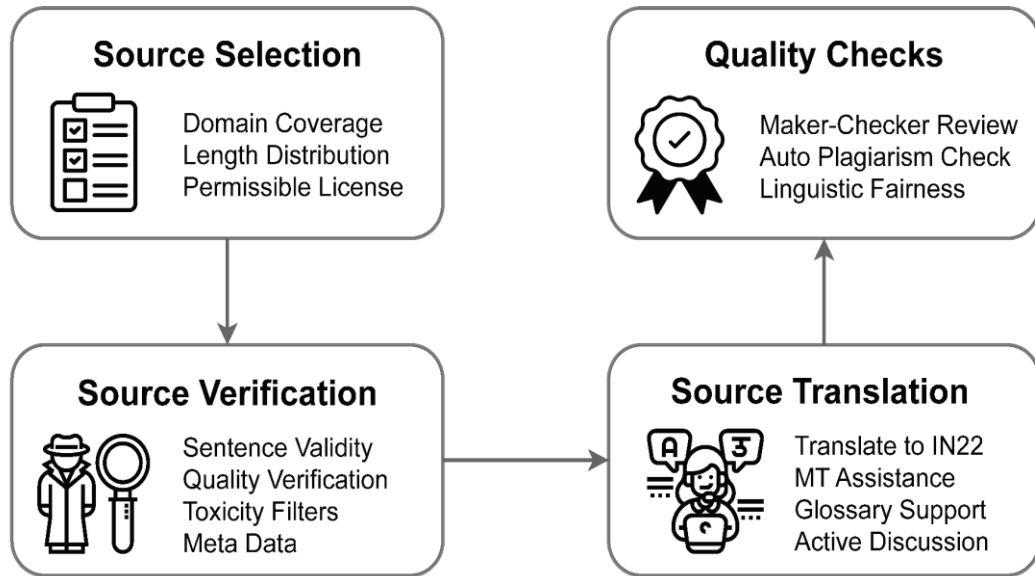
*Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. Language-agnostic BERT sentence embedding, ACL 2022.

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

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



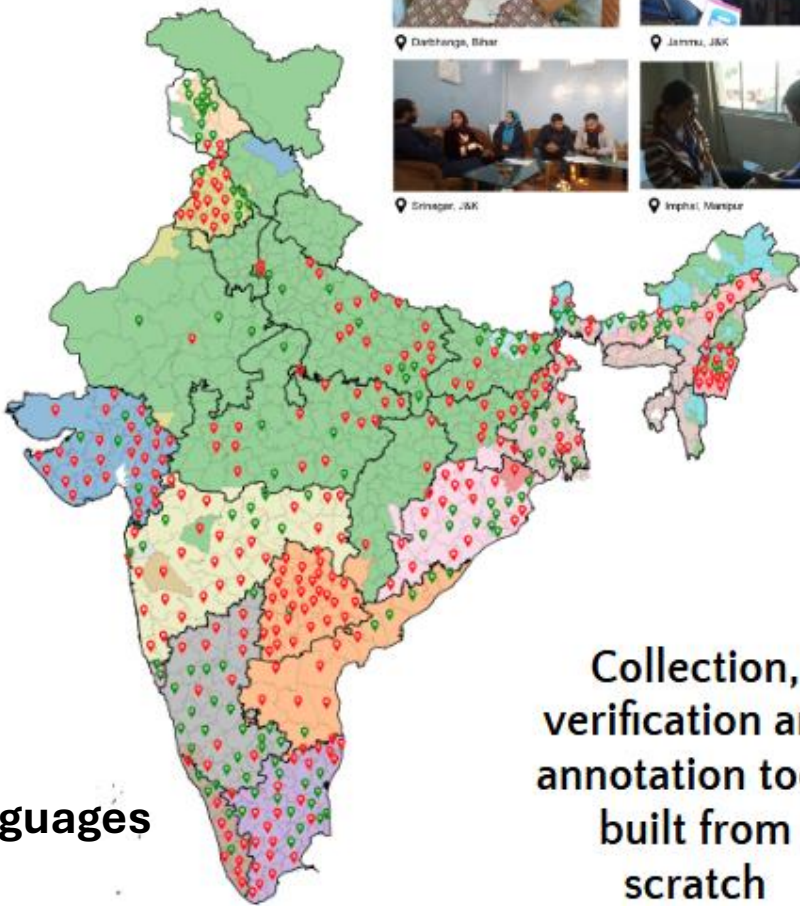
- Need processes in place to ensure high quality
- Provide tools to make translators productive

The screenshot shows the Shoonya web application interface for translation tasks. The top navigation bar includes links for Organization, Projects, Datasets, Analytics, and Admin, along with a user profile for Ishvinder. The main content area displays a task titled "#2054854" by Ishvinder Sethi. It features a "Draft" status and a "Next" button. The task details include a "Source sentence" in English, an "Assamese translation" in Assamese script, and a "Machine translation" in Assamese script. A "Context" section provides background information on the Nilamata Purana. The interface also includes a "Skip" button and an "Update" button.

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+ 1893
Districts Personnels
12000
HOURS
COLLECTED



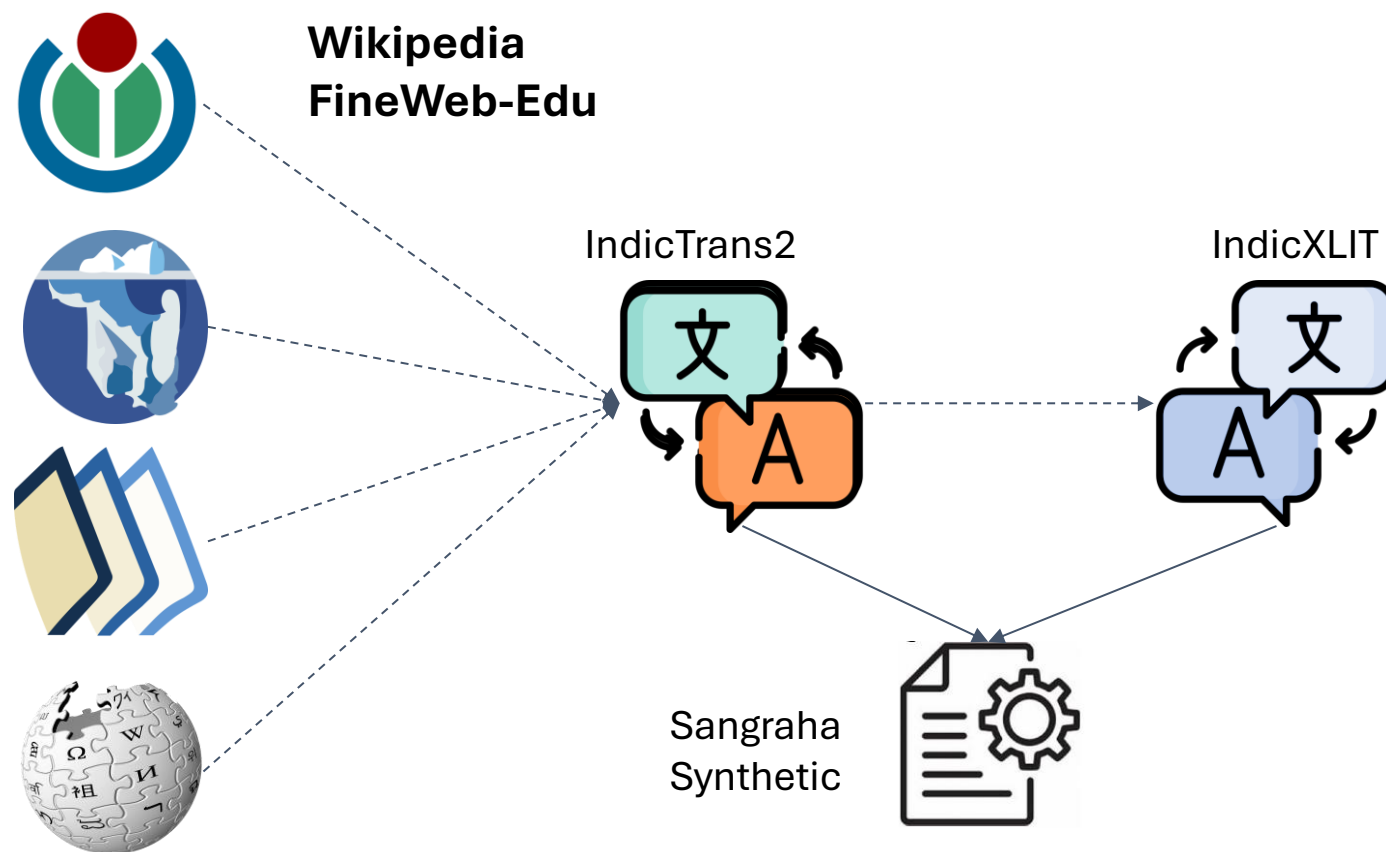
Generate Multilingual Data

*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

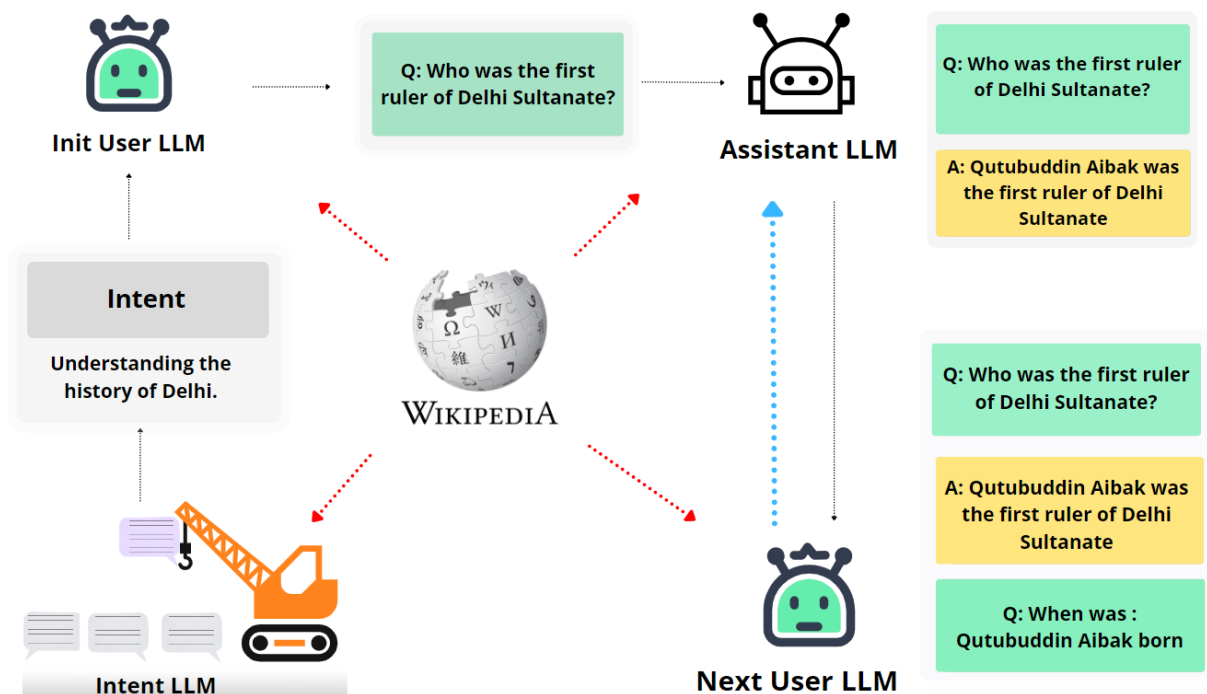


Creativity is
the limit

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



Wiki-Chat

Can you tell me more about the Lotus Temple? What is it known for and what makes it unique?

The Lotus Temple is a Bahá'í House of Worship located in South Delhi, India. It is known for its distinctive lotus-like architecture, which is made up of 27 marble petals arranged in clusters. The temple has received a wide range of attention in various venues for its exceptional beauty and architectural design. It can accommodate up to 2,500 people and is open to people of all religions, who come to pray, meditate, or contemplate. The temple's unique design and serene atmosphere make it a popular destination for tourists and locals alike. It is surrounded by nine pools of water and gardens, adding to its peaceful and picturesque setting. The Lotus Temple is one of the most visited buildings in the world, attracting millions of visitors every year.

Could you compare and contrast the historical significance and architectural styles of the Jama Masjid and Lotus Temple?

Sure, I'd be happy to help with that.

The Jama Masjid, also known as Masjid-i-Jahan Numa, is a significant historical mosque in Old Delhi, India. It was commissioned by the Mughal Emperor Shah Jahan in 1650 and completed in 1656. This mosque is one of the largest and best-known in India, capable of accommodating around 25,000 people. Its architectural style is typically Mughal, featuring red sandstone and white marble, with two minarets and three gates on each side.

The Lotus Temple, located in South Delhi, is a Bahá'í House of Worship. Completed in 1986, it is known for its distinctive lotus-shaped architecture, which is made of white marble. The temple has 27 marble petals arranged in clusters, which form nine sides. The architectural style is modern and unique, representing the Bahá'í faith's principle of the unity of all religions.

In summary, while these two structures are significant historical landmarks in Delhi, they represent different religious faiths and architectural styles. The Jama Masjid is a classic example of Mughal architecture, the Lotus Temple is a modern representation of the Bahá'í faith's unity of religions.

(a) English

Wiki-Chat

లోటస్ టెంపుల్ గురించి మరింత చెప్పగలరా? ఇది దేనికి ప్రసిద్ధి చెందింది మరియు దాని ప్రత్యేకత ఏమిటి?

లోటస్ టెంపుల్ అనేది భారతదేశంలోని దక్షిణ ఢిల్లీలో ఉన్న బహాయ్ ఆరాధన గృహం. ఇది దాని విలక్షణమైన కమలం-వంటి నిర్మాణానికి ప్రసిద్ధి చెందింది. ఇది సమూహాలలో అమర్చబడిన 27 పాలరాతి రేకులతో రూపొందించబడింది. ఈ ఆలయం అసాధారణమైన అందం మరియు నిర్మాణ రూపకల్పన కోసం వివిధ వేదికలలో విస్తృత దృష్టిని పొందింది. ఇది గరిష్టంగా 2,500 మంది వ్యక్తులకు వసతి కల్పిస్తుంది మరియు ప్రార్థన చేయడానికి, ధ్యానం చేయడానికి లేదా ధ్యానించడానికి వచ్చే అన్ని మతాల ప్రజలకు అందుబాటులో ఉంటుంది. ఆలయం యొక్క ప్రత్యేకమైన డిజైన్ మరియు నిర్మలమైన వాతావరణం దీనిని పర్యాటకులకు మరియు స్థానికులకు ఒక ప్రసిద్ధ గమ్యస్థానంగా మార్చింది. దాని చుట్టూ తొమ్మిది నీటి కిలుసులు మరియు తోటలు ఉన్నాయి, దాని ప్రశాంతమైన మరియు సుందరమైన అమరికను జోడించడం. లోటస్ టెంపుల్ ప్రపంచంలో అత్యధికంగా సందర్శించే భవనాలలో ఒకటి. ప్రతి సంవత్సరం మిలియన్ల మంది సందర్శకులను ఆకర్షిస్తుంది.

మీరు జామా మసీదు మరియు లోటస్ టెంపుల్ యొక్క చారిత్రక ప్రాముఖ్యత మరియు నిర్మాణ శైలులను పోల్చి చూడగలరా?

ఖచ్చితంగా, నేను దానిలో సహాయం చేయడానికి సంతోషిస్తున్నాను.

జామా మసీదును మస్జిద్-ఇ-జహాన్ నుండి అని కూడా పిలుస్తారు. ఇది భారతదేశంలోని పాత ఢిల్లీలో ఉన్న ఒక ముఖ్యమైన చారిత్రక మసీదు. ఇది 1650లో మొఘల్ చక్రవర్తి షాజహాన్ చేత ప్రారంభించబడింది మరియు 1656లో పూర్తయింది. ఈ మసీదు దాదాపు 25,000 మందికి వసతి కల్పించే సామర్థ్యం ఉన్న భారతదేశంలోని అతిపెద్ద మరియు ప్రసిద్ధి చెందిన వాటిలో ఒకటి. దీని నిర్మాణ శైలి సాధారణంగా మొఘల్, పర్ ఇసుకరాయి మరియు తల్లిన పాలరాయిని కలిగి ఉంటుంది. ప్రతి వైపు రెండు మినార్లు మరియు మూడు గేట్లు ఉంటాయి.

దక్షిణ ఢిల్లీలో ఉన్న లోటస్ టెంపుల్, బహాయ్ ఆరాధన గృహం, 1986లో పూర్తయింది. ఇది తల్లిన పాలరాతితో తయారు చేయబడిన విలక్షణమైన తామర ఆకారపు నిర్మాణానికి ప్రసిద్ధి చెందింది. ఆలయంలో తొమ్మిది వైపులా ఉండే 27 పాలరాతి రేకులు గుత్తులుగా అమర్చబడి ఉన్నాయి. నిర్మాణ శైలి ఆధునికమైనది మరియు విశిష్టమైనది, అన్ని మతాల వ్యక్తత యొక్క బహాయ్ విశ్వాసం యొక్క స్ఫూర్తిని సూచిస్తుంది.

సారాంశంలో, ఈ రెండు నిర్మాణాలు ఢిల్లీలో ముఖ్యమైన చారిత్రక మైలురాయి అయితే, అవి విభిన్న మత విశ్వాసాల మరియు నిర్మాణ శైలులను సూచిస్తాయి. జామా మసీదు మొఘల్ వాస్తవశైలికి ఒక అద్భుతమైన ఉదాహరణ, లోటస్ టెంపుల్ బహాయ్ విశ్వాసం యొక్క మతాల వ్యక్తత ఆధునిక ప్రాతినిధ్యం.

(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

NLP Infrastructure: Raw corpora



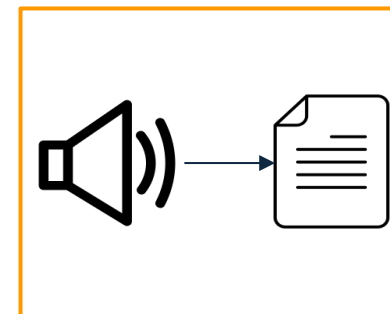
IndicCorp

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)

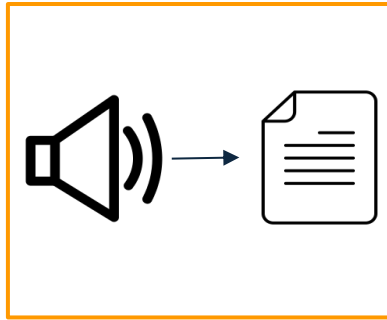
Compact pre-trained models for NLU & NLG

Data for various foundational tasks



BPCCC

Parallel corpus,
translation models
between English & 22
Indic languages



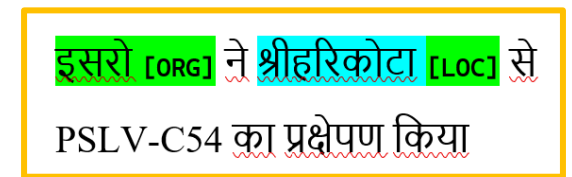
Shrutilipi, IndicVoices & KathBath

ASR datasets for 22
Indian languages



Aksharantar

Transliteration datasets
for 20 Indic languages



Naamapadam

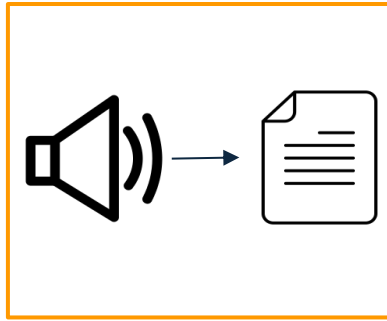
Datasets and models for
Named Entity
Recognition in 11 Indian
languages

Models for foundational tasks



IndicTransv2

Parallel corpus,
translation models
between English & 22
Indic languages



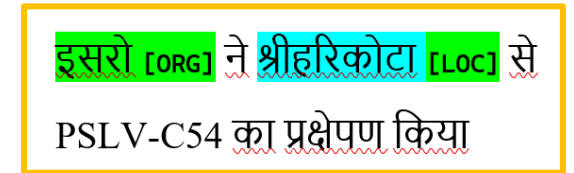
IndicConformer

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 →](#)

Chitrlekha

Chitrlekha 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

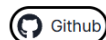
<https://ai4bharat.iitm.ac.in>

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), Ol Chiki (Santali), Meitei (Manipuri), Latin (English), and Devanagari (used for all the remaining languages).



Select Service: *

ai4bharat/indictrans--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>



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...



Audio Transcription (NEW)
Try out our Indic ASR Arena



Text to Speech (NEW)
Try out our Indic TTS Arena



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://anoopkunchukuttan.github.io>

<http://huggingface.co/Al4Bharat>

Acknowledgments: All my collaborators, colleagues and students at Al4Bharat