

Building Multilingual NLP Datasets at scale

Anoop Kunchukuttan

Microsoft



AI4Bharat



IASNLP, IIIT Hyderabad, 23 June 2024

Deep Learning has led to tremendous success

Transformer-based self-supervised pre-trained models

GLUE Benchmark

Rank	Name	Model	URL	Score
1	ERNIE Team - Baidu	ERNIE	🔗	90.9
2	DeBERTa Team - Microsoft	DeBERTa / TuringNLRv4	🔗	90.8
3	HFL iFLYTEK	MacALBERT + DKM		90.7
+ 4	Alibaba DAMO NLP	StructBERT + TAPT	🔗	90.6
+ 5	PING-AN Omni-Sinitic	ALBERT + DAAF + NAS		90.6
6	T5 Team - Google	T5	🔗	90.3
7	Microsoft D365 AI & MSR AI & GATECHMT-DNN-SMART		🔗	89.9
+ 8	Huawei Noah's Ark Lab	NEZHA-Large		89.8
+ 9	Zihang Dai	Funnel-Transformer (Ensemble B10-10-10H1024)	🔗	89.7

SUPERGLUE Benchmark

Rank	Name	Model	URL	Score
+ 1	Zirui Wang	T5 + Meena, Single Model (Meena Team - Google Brain)		90.4
+ 2	DeBERTa Team - Microsoft	DeBERTa / TuringNLRv4	🔗	90.3
3	SuperGLUE Human Baselines	SuperGLUE Human Baselines	🔗	89.8
+ 4	T5 Team - Google	T5	🔗	89.3
+ 5	Huawei Noah's Ark Lab	NEZHA-Plus	🔗	86.7
+ 6	Alibaba PAI&ICBU	PAI Albert		86.1

Tasks: Classification, Structured Prediction, Span Prediction, Semantic Similarity

Maths

Prompt: Can you write a proof that there are infinitely many primes, with every line that rhymes?

GPT-4:

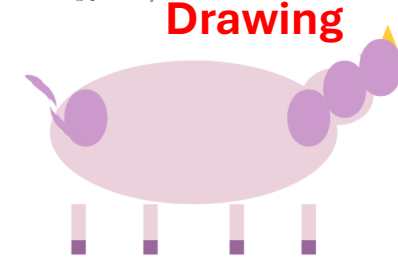
Yes, I think I can,
Though it might take a clever plan.
I'll start by noting Euclid's proof,
Which shows that primes aren't just aloof.
Assume we have a finite list,
Of primes, and that none have been missed.
Multiply them all together,
And add one, just to be clever.
The result will be a number,
That has no prime factors, oh wonder!
But every number has a prime,
That divides it, it's just a matter of time.
So we've found a contradiction,
And our finite list needs eviction.
There must be infinitely many primes,
And that's the end of my rhyming lines.s

Prompt: Can you generate a pyplot for the following data: $x = [1, 3, 5, 6, 8]$, $y1 = [2, 3, 5, 18, 1]$, $y2 = [3, 5, 6, 8, 1]$, $y3 = [5, 1, 2, 3, 4]$, $y4 = [9, 7, 2, 3, 1]$. I want $y1, y2$ to be in the same plot, but $y3$ is in another plot next to that plot, $y4$ is in below. I want the legend of $y1$ to be "bob", $y2$ to be "alice", $y3$ to be "bilbo", $y4$ to be "allie". I want the x -axis to be labeled with "time" and y axis to be labeled with "money". I want to add a 10 %-40% random error bar to each curve, through all times (including non-integers). I want smoothed curves to show the plot, and smoothed error bar. Do not use linear interpolation, use smooth interpolation! I want to also add some small zig-zag to the smoothed curve to make it look more real. I want to put a baseline as the mean of each line. I want to put a pie chart below indicating the fraction of the four people in each time step. I also want an animation to show how the fractions are changing in the pie chart in **continuous time**. Interpolate the missing fractions! I also want vertical line animation in other three plots to match the pie chart. I want the fanciest plot. Please add as many fancy things as possible.

Here is a possible code to generate a pyplot for the following data, with some comments to explain the steps:

Prompt: Draw a unicorn in TikZ.

GPT-4: [Produces \LaTeX compiling to following picture.]

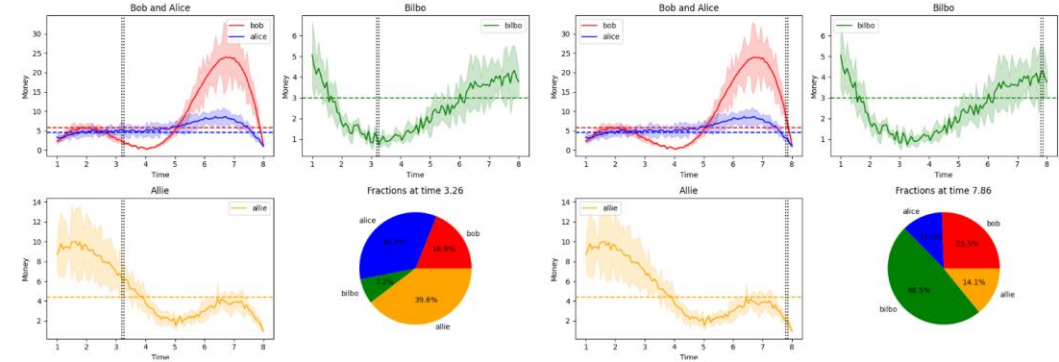


Plotting

Prompt: Can you generate a pyplot for the following data: $x = [1, 3, 5, 6, 8]$, $y1 = [2, 3, 5, 18, 1]$, $y2 = [3, 5, 6, 8, 1]$, $y3 = [5, 1, 2, 3, 4]$, $y4 = [9, 7, 2, 3, 1]$. I want $y1, y2$ to be in the same plot, but $y3$ is in another plot next to that plot, $y4$ is in below. I want the legend of $y1$ to be "bob", $y2$ to be "alice", $y3$ to be "bilbo", $y4$ to be "allie". I want the x -axis to be labeled with "time" and y axis to be labeled with "money". I want to add a 10 %-40% random error bar to each curve, through all times (including non-integers). I want smoothed curves to show the plot, and smoothed error bar. Do not use linear interpolation, use smooth interpolation! I want to also add some small zig-zag to the smoothed curve to make it look more real. I want to put a baseline as the mean of each line. I want to put a pie chart below indicating the fraction of the four people in each time step. I also want an animation to show how the fractions are changing in the pie chart in **continuous time**. Interpolate the missing fractions! I also want vertical line animation in other three plots to match the pie chart. I want the fanciest plot. Please add as many fancy things as possible.

Here is a possible code to generate a pyplot for the following data, with some comments to explain the steps:

GPT-4: [We give two snapshots from the animation resulting from the produced code]



Explanations

GPT-4:

This is a conversation between a user and an artificial intelligence agent that follows instructions or answers questions. The AI always follows the instructions correctly.

User: Can you compose a short tune (say four to eight bars) using ABC notation?

AI: There are many possible ways to compose a short tune in abc notation, but here is one example:

X:1

T:Example Tune

M:4/4

L:1/8

K:C

|:G2AB c2BA | G2AB c2c2 | edcB c2BA | G4-G2z2 :|

|:c2de f2ed | c2de f2f2 | g2fe d2cB | c4-c2z2 :|

This tune has two four-bar phrases, each repeated once, forming an A part and a B part. The key is C major.

GPT-4:

User: After the first bar G2AB c2BA, you chose G to start the next bar. Give an example of a change to the notes in the first bar that would have caused you to pick D instead of G as the first note in the second bar.

AI:One possible change to the notes in the first bar that would have caused me to pick D instead of G as the first note in the second bar is:

|:G2AB c2Bc | ...

This change creates a stronger cadence on the C note at the end of the first bar, which is the fifth degree of the scale. This suggests a modulation to the dominant key of G major, which has D as its fifth degree. Therefore, starting the second bar with D would create a smooth transition to the new key and imply a harmonic progression of V-I.

Prompt:

```
struct mystruct_A { char a; int b; char c; } x;
struct mystruct_B { int b; char a; char c; } y;

#include <stdio.h>
int main(void) {
    printf("Size of x: %zu bytes\n", sizeof(x));
    printf("Size of y: %zu bytes\n", sizeof(y));
    return 0;
}
```

Possible Output:

GPT-4:

There are different ways to implement and align structures in memory, so the exact output may vary depending on the compiler, platform, and options used. However, one possible output is:

Size of x: 12 bytes
Size of y: 8 bytes

This output assumes that the compiler uses 4-byte alignment for structures, meaning that each structure member starts at an offset that is a multiple of 4 bytes from the beginning of the structure. This may result in some padding bytes being inserted between or after the members to maintain alignment. For example, in x, there may be 3 padding bytes after a and 3 padding bytes after c, while in y, there may be no padding at all. The size of a structure is the sum of the sizes of its members plus any padding bytes.

Understanding Programs

Large Language Models show great performance on diverse open-ended tasks

Transformer-based self-supervised decoder only models

Model	Arena Score
Gemini-2.5-Pro-Exp-03-25	1439
o3-2025-04-16	1418
ChatGPT-4o-latest (2025-03-26)	1408
chocolate (Early Grok-3)	1402
Grok-3-Preview-02-24	1402
Llama-4-Maverick-03-26-Experimental	1401
GPT-4.5-Preview	1398
Gemini-2.5-Flash-Preview-04-17	1393
Gemini-2.0-Flash-Thinking-Exp-01-21	1380
Gemini-2.0-Pro-Exp-02-05	1380
ChatGPT-4o-latest (2025-	1374

(LMSys Chatbot Arena)

	Qwen3-235B-A22B MoE	Qwen3-32B Dense	OpenAI-o1 2024-12-17	Deepseek-R1	Grok 3 Beta Think	Gemini2.5-Pro	OpenAI-o3-mini Medium
ArenaHard	95.6	93.8	92.1	93.2	-	96.4	89.0
AIME'24	85.7	81.4	74.3	79.8	83.9	92.0	79.6
AIME'25	81.5	72.9	79.2	70.0	77.3	86.7	74.8
LiveCodeBench v5, 2024.10-2025.02	70.7	65.7	63.9	64.3	70.6	70.4	66.3
CodeForces Elo Rating	2056	1977	1891	2029	-	2001	2036
Aider Pass@2	61.8	50.2	61.7	56.9	53.3	72.9	53.8
LiveBench 2024-11-25	77.1	74.9	75.7	71.6	-	82.4	70.0
BFCL v3	70.8	70.3	67.8	56.9	-	62.9	64.6
Multif 8 Languages	71.9	73.0	48.8	67.7	-	77.8	48.4

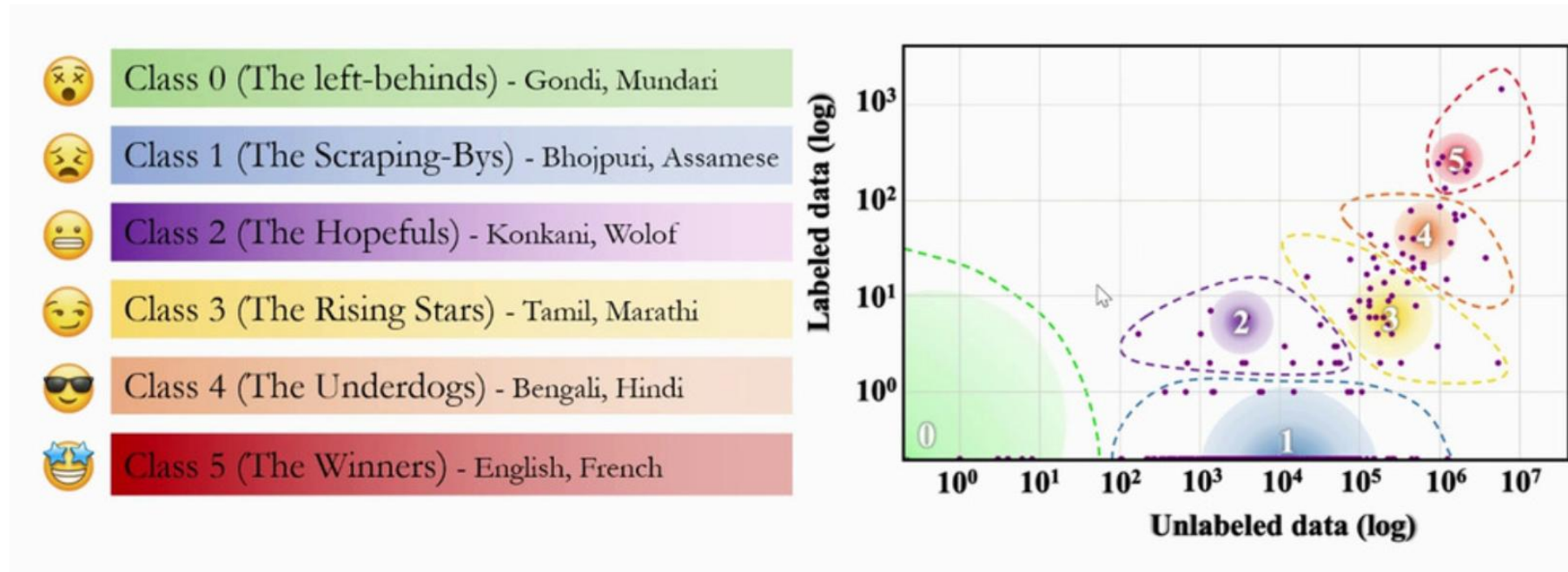
Compilation of tasks requiring reasoning skills

(Qwen 3)

Tasks: Open-ended Question Answering evaluated on dynamic questions based on human preferences

Disparity in linguistic resources has always been an issue for NLP

Wikipedia/CommonCrawl data as a proxy for monolingual data availability



How do we bring the state-of-the-art NLP solutions to all languages?

Can we train such large models for all languages?

Benefits of LLMs are mostly limited to English

Language	Cat.	ChatGPT	
		(en)	(spc)
English	H	70.2	70.2
Russian	H	60.8	45.4
German	H	64.5	51.1
Chinese	H	58.2	35.5
French	H	64.8	42.2
Spanish	H	65.8	47.4
Vietnamese	H	55.4	44.8
Turkish	M	57.1	37.1
Arabic	M	55.3	22.3
Greek	M	55.9	54.5
Thai	M	44.7	11.5
Bulgarian	M	59.7	44.6
Hindi	M	48.8	5.6
Urdu	L	43.7	6.3
Swahili	X	50.3	40.8

Results on XNLI

Language	Cat.	ChatGPT(en)	
		EM	F1
English	H	56.0	74.9
Russian	H	30.2	49.1
German	H	45.9	65.8
Chinese	H	37.1	42.3
Spanish	H	41.8	65.8
Vietnamese	H	36.1	57.3
Turkish	M	34.5	56.4
Arabic	M	32.0	50.3
Greek	M	29.7	45.0
Thai	M	31.2	43.4
Hindi	M	17.5	37.8
Average		35.6	53.5

Results on QnA

	#langs.	avg. chrF	avg. BLEU
ChatGPT (0-shot)	203	32.3	16.7
ChatGPT (5-shot)	203	33.1	17.3
GPT-4	20	44.6	24.6
NLLB	201	45.3	27.1
Google	115	52.2	34.6

**Performance on translation
averaged across languages**

Lang.	ChatGPT		NLLB	
	BLEU	chrF	BLEU	chrF
srp_Cyrl	1.36	3.26	43.4	59.7
kon_Latn	0.94	8.50	18.9	45.3
tso_Latn	2.92	15.0	26.7	50.0
kac_Latn	0.04	2.95	14.3	37.5
nso_Latn	3.69	16.7	26.5	50.8
jpn_Jpan	28.4	32.9	20.1	27.9
nno_Latn	37.1	58.7	33.4	53.6
zho_Hans	36.3	31.0	26.6	22.8
zho_Hant	26.0	24.4	12.4	14.0
acm_Arab	28.2	44.7	11.8	31.9

**Performance on translation
High vs low resource**

- Significant gap between English and other languages on multiple tasks
- High-resource and Latin script languages can give good performance on GPT
- Poor performance on low-resource languages
- Translate-test is a strong baseline
- Open-source models lag behind GPT models → they are very English heavy

(BUFFET, MEGA, ChatGptMT, ChatGptMLing)

We are faced with a huge data skew

Raw Text Corpora	Wikipedia articles	English	6m
		Hindi	
Parallel Corpora	Sentence pairs	En-fr (OPUS)	150k
		En-hi (IITB)	500m
NER Corpora	Tokens	en (CoNLL 2003)	200k
		hi (FIRE)	40k
QA	Question-Answer Pairs	en (SQuAD 1.1)	100k
		hi (MMQA)	4.6k

How do we address this data skew at scale?

Mining Datasets

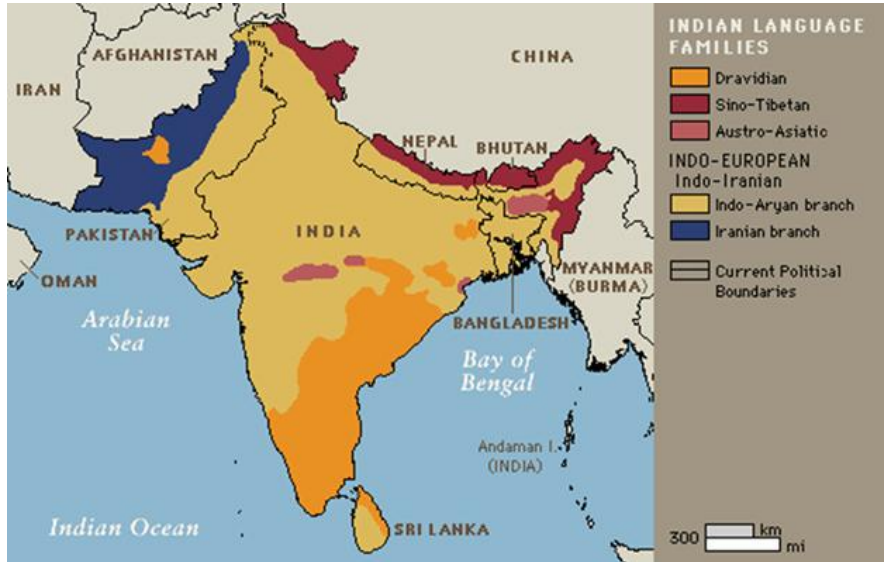
Synthesizing Datasets

Manually Constructing Datasets

We will discuss in the context of Indian languages and our experiences in that respect

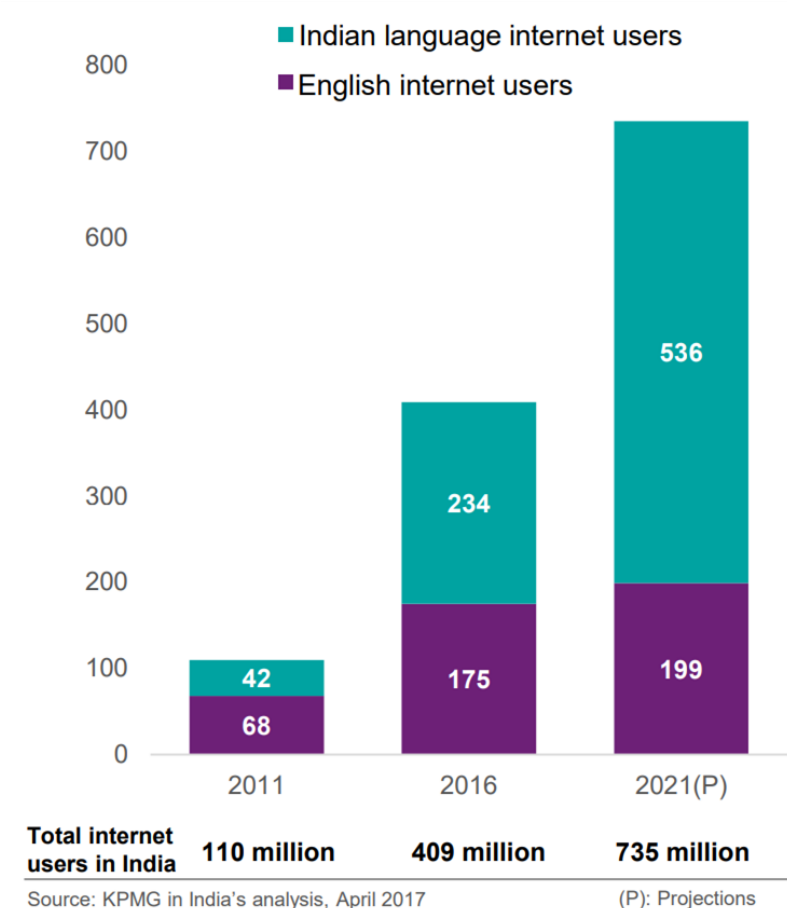
Lot of other languages and groups have explored similar approaches

Usage and Diversity of Indian Languages



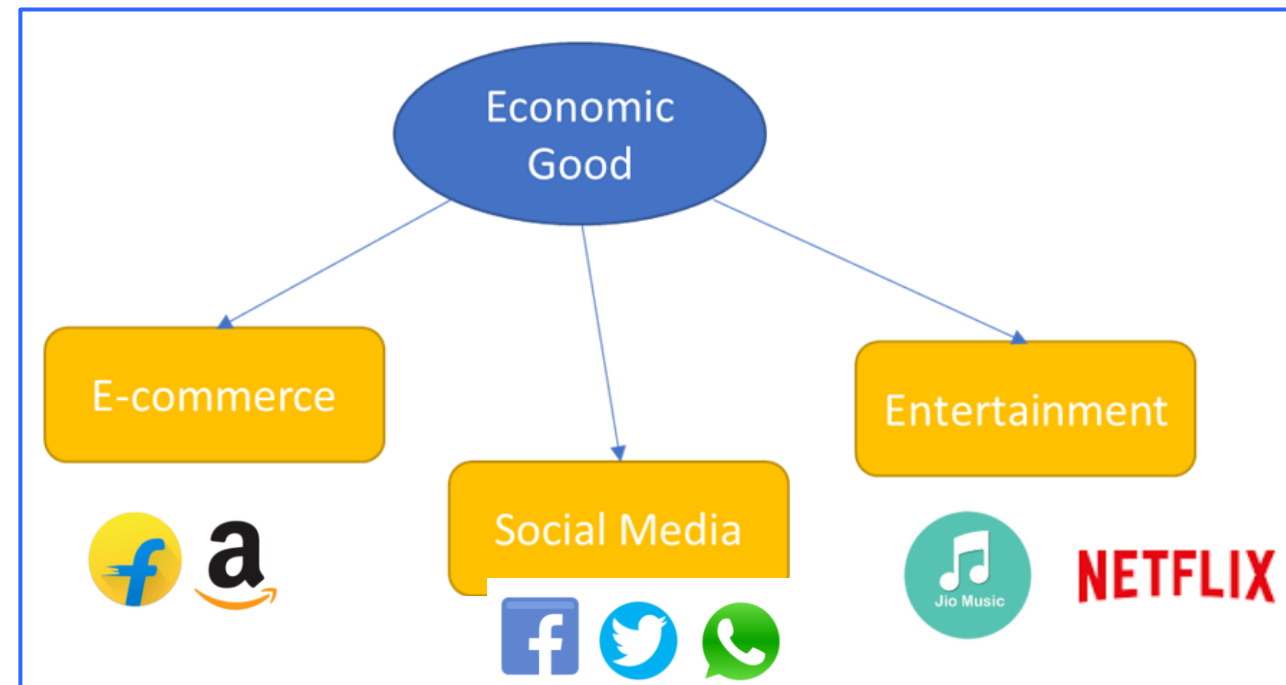
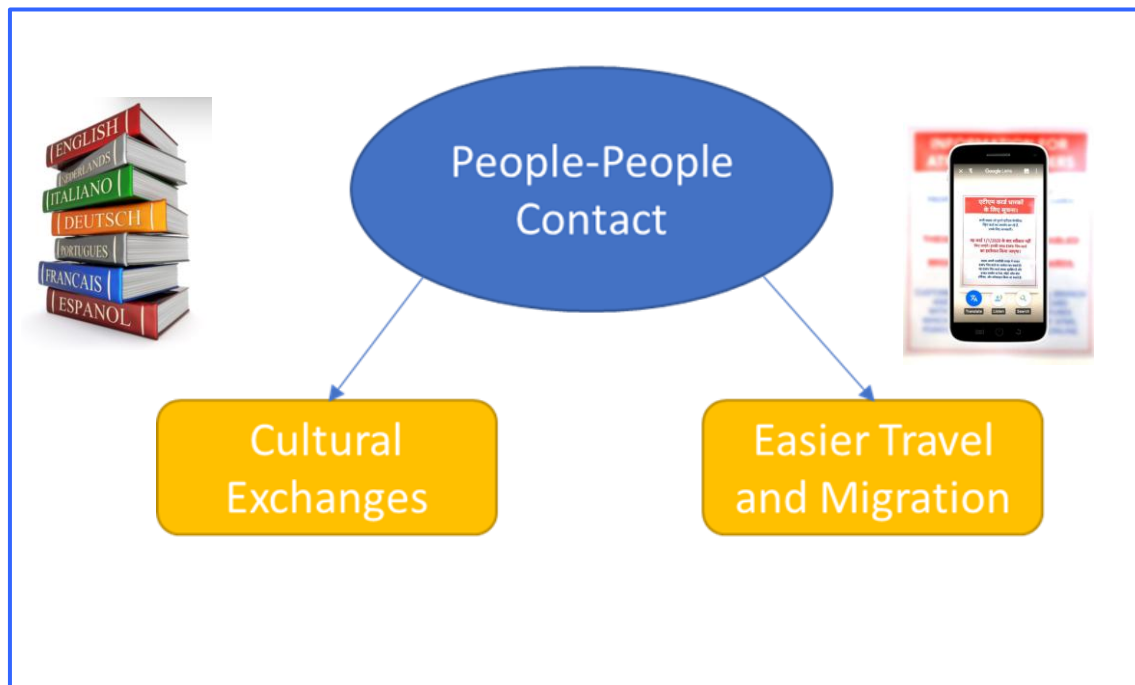
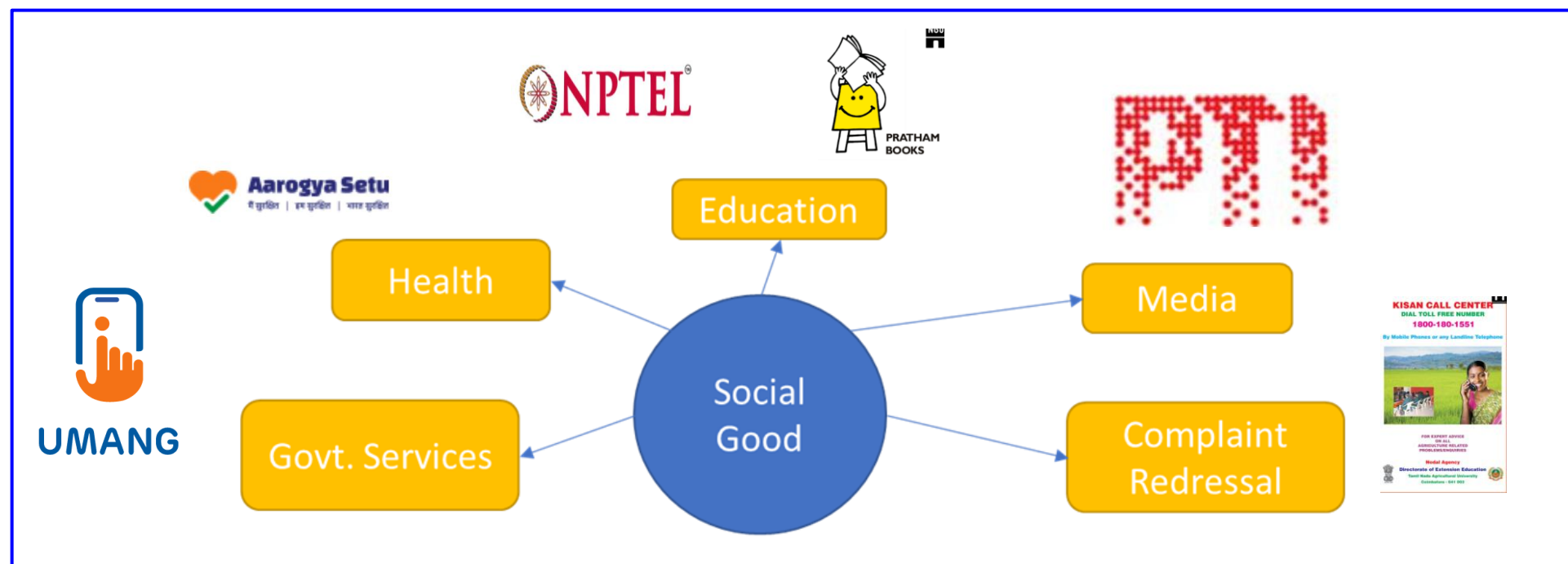
- 4 major language families
- 22 scheduled languages
- 125 million English speakers
- 8 languages in the world's top 20 languages
- 30 languages with more than 1 million speakers

Sources: Wikipedia, Census of India 2011

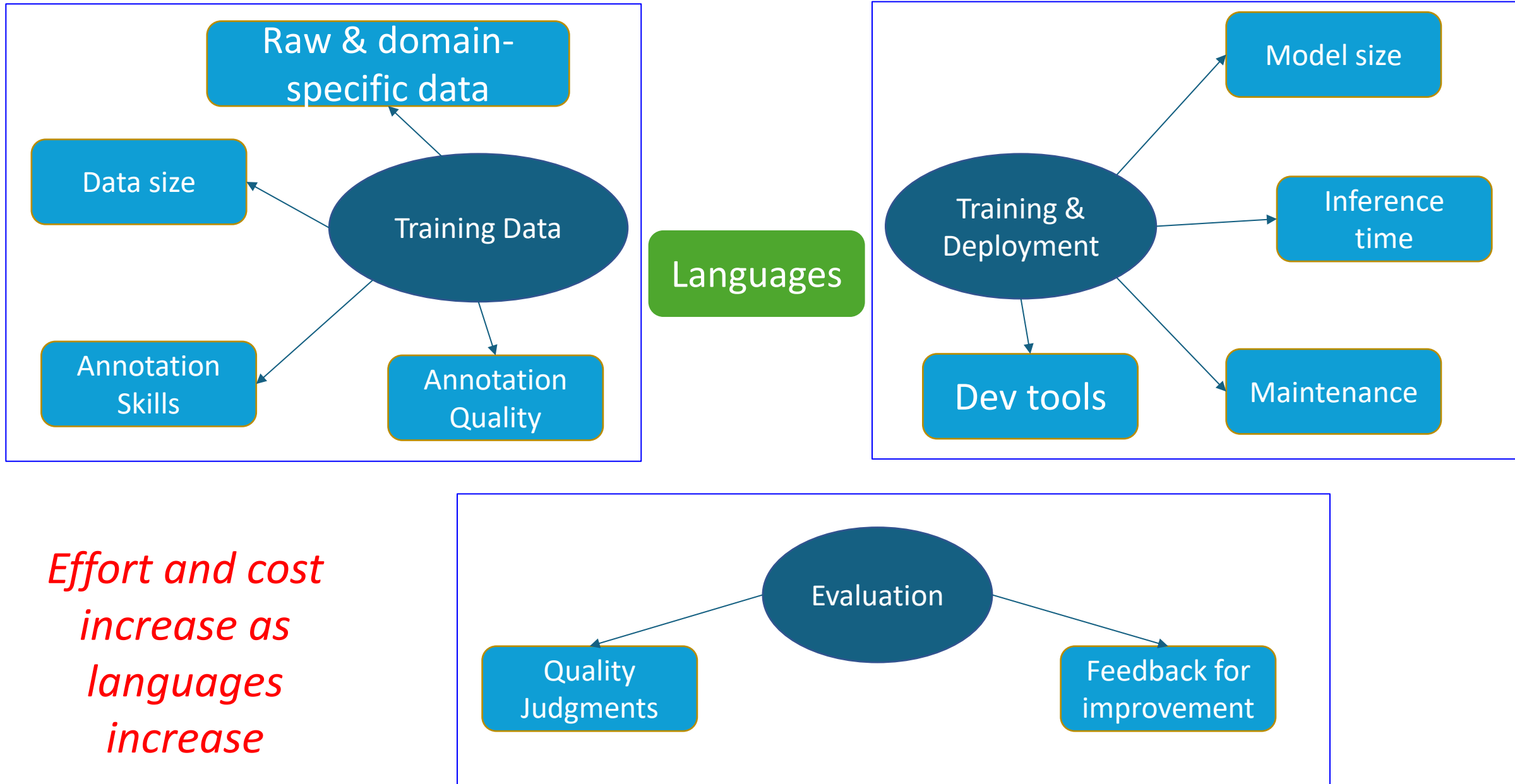


Internet User Base in India (in million)

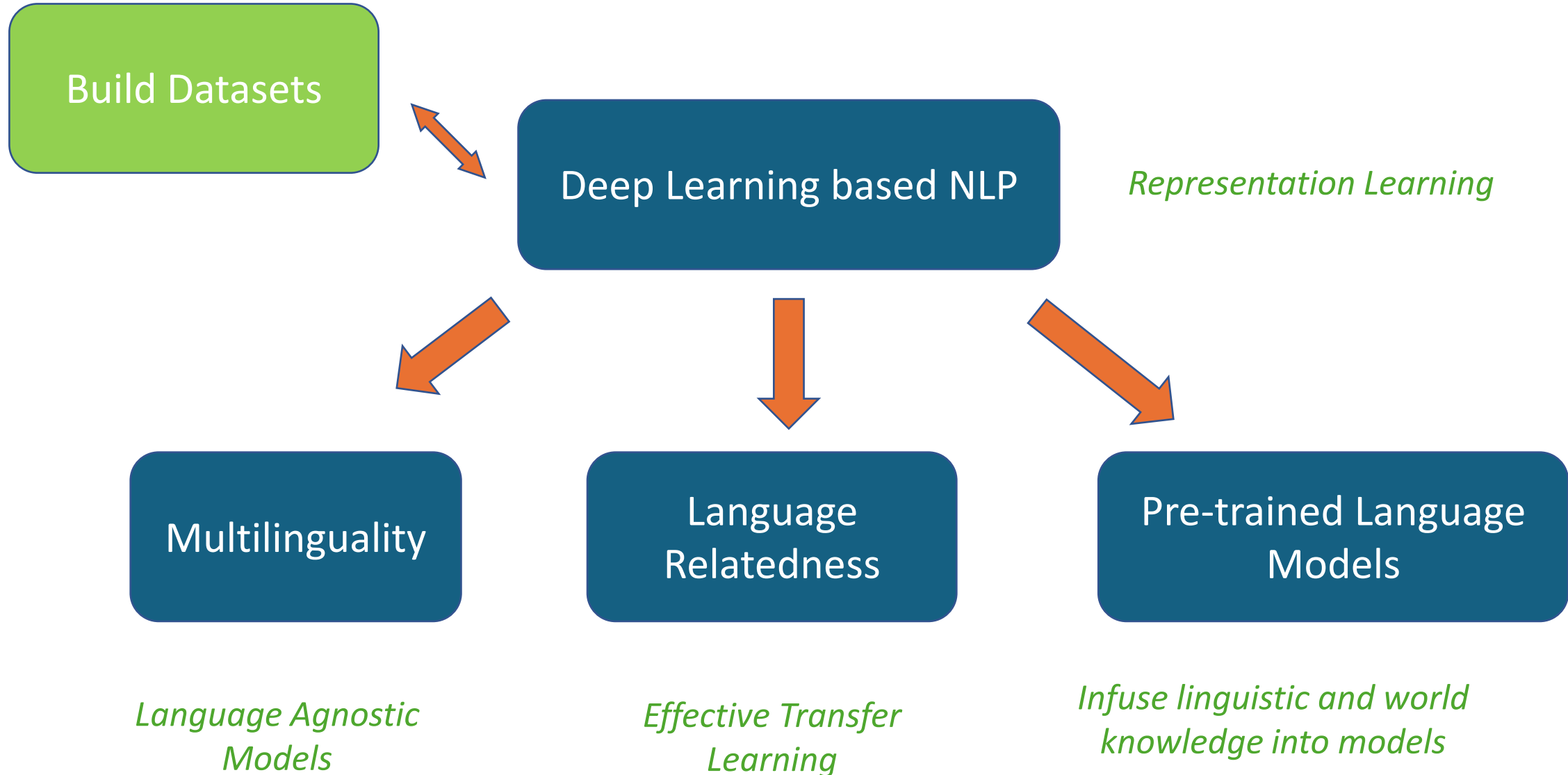
Source: Indian Languages:
Defining India's Internet KPMG-Google Report 2017



Scalability Challenges for NLP solutions



The Opportunity for low-resource multilingual NLP



The proposed recipe for multilingual NLP



CORPUS ... the open parallel corpus



mBERT
mBART/mT5
AYA

Collect a large amount
of monolingual and/or
parallel data

Lot of monolingual data

Learn a **multilingual language
model**

**Lot of memory &
computation**

But not all languages have such data

Joint learning & Data Curation is the Key

Hi	विश्वविद्यालय जहाँ स्थित है, उसके पूर्वी दिशा में बने परिसरों को क्या नाम दिया गया है?
जब 1919 में यूसीएलए ने अपना नया परिसर खोला, तब इसमें चार इमारतें थी। [...] परिसर अनौपचारिक रूप से उत्तरी परिसर और दक्षिणी परिसर में विभाजित है, जो दोनों विश्वविद्यालय की जमीन के पूर्वी हिस्से में स्थित हैं। [...] दक्षिणी परिसर में भौतिक विज्ञान, जीव विज्ञान, इंजीनियरिंग, मनोविज्ञान, गणितीय विज्ञान, सभी स्वास्थ्य से संबंधित क्षेत्र और यूएलसीए मेडिकल सेंटर स्थित है।	

*Finetune the language model
on reasonable amount of
data from one or more
languages*

- Task-specific
- General instruction-response
- Human Preference

**Supervised data limited
for many languages**

Multilingual Data

- Raw Text Corpora
- Cross-lingual Corpora
 - Machine Translation Corpora
 - Machine Transliteration Corpora
- Mining Task data/Instruction data
- Synthesizing Task data/Instruction data
- Multilingual, Multimodal data

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*

*What properties do we
want to see in
multilingual corpora?*

Large-scale, Document-level Datasets

High Quality Documents

Wide coverage of topics

Representation of culture-specific data, native literature

Capture data in different modalities and genres

Data to Help Cross-lingual transfer with English



Large-scale, Document-level Datasets



High Quality Documents



Wide coverage of topics

Multilingual corpora like mC4, CC-100, CulturaX are good starting points



Representation of culture-specific data, native literature



Capture data in different modalities and genres



Data to Help Cross-lingual transfer with English

Build custom language (group) specific collections to address gaps

Raw Text Data Collection

IndicCorp v1

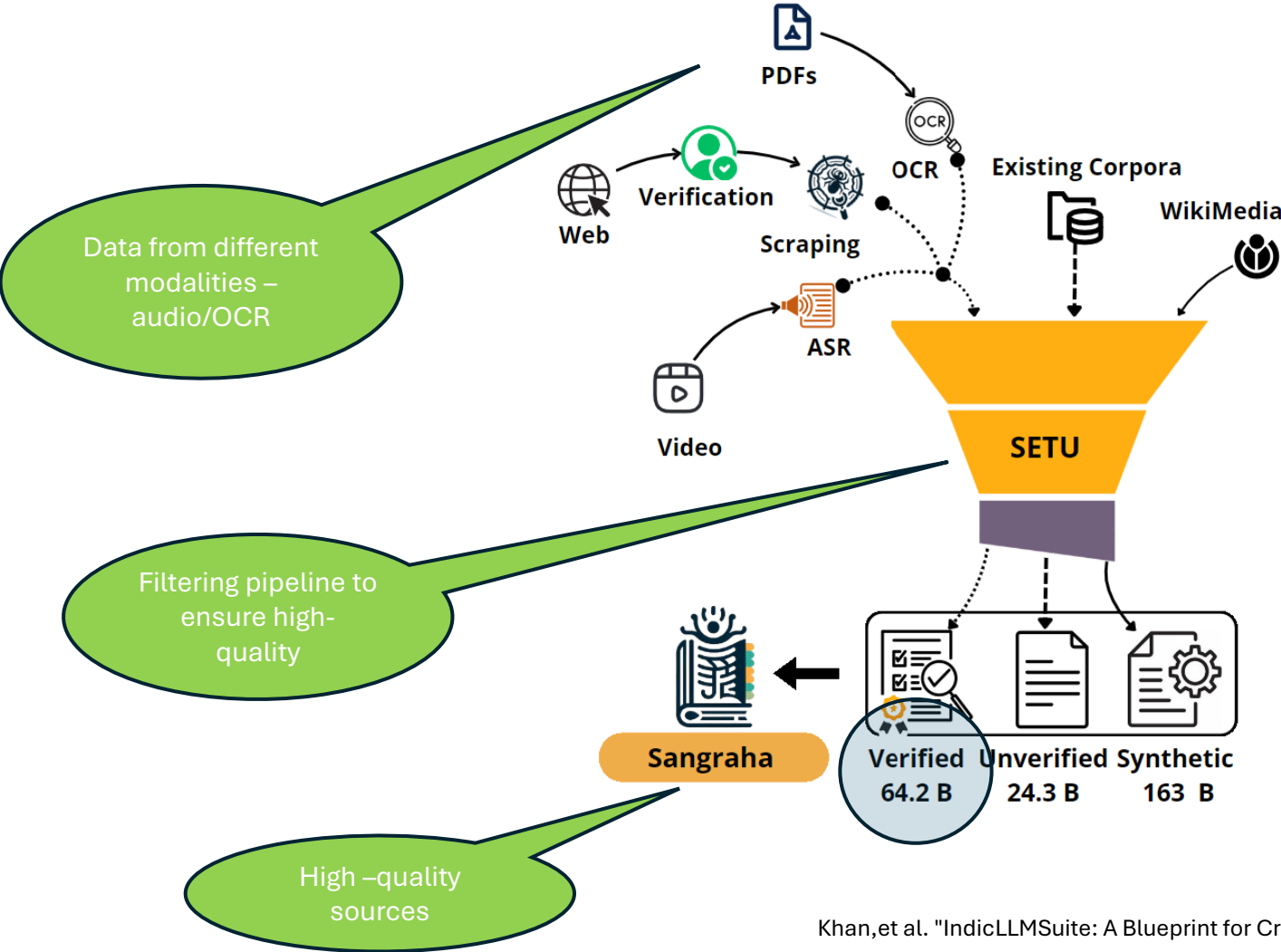
Sentence-level
Web-sources

IndicCorp v2

Larger corpora
Larger language coverage

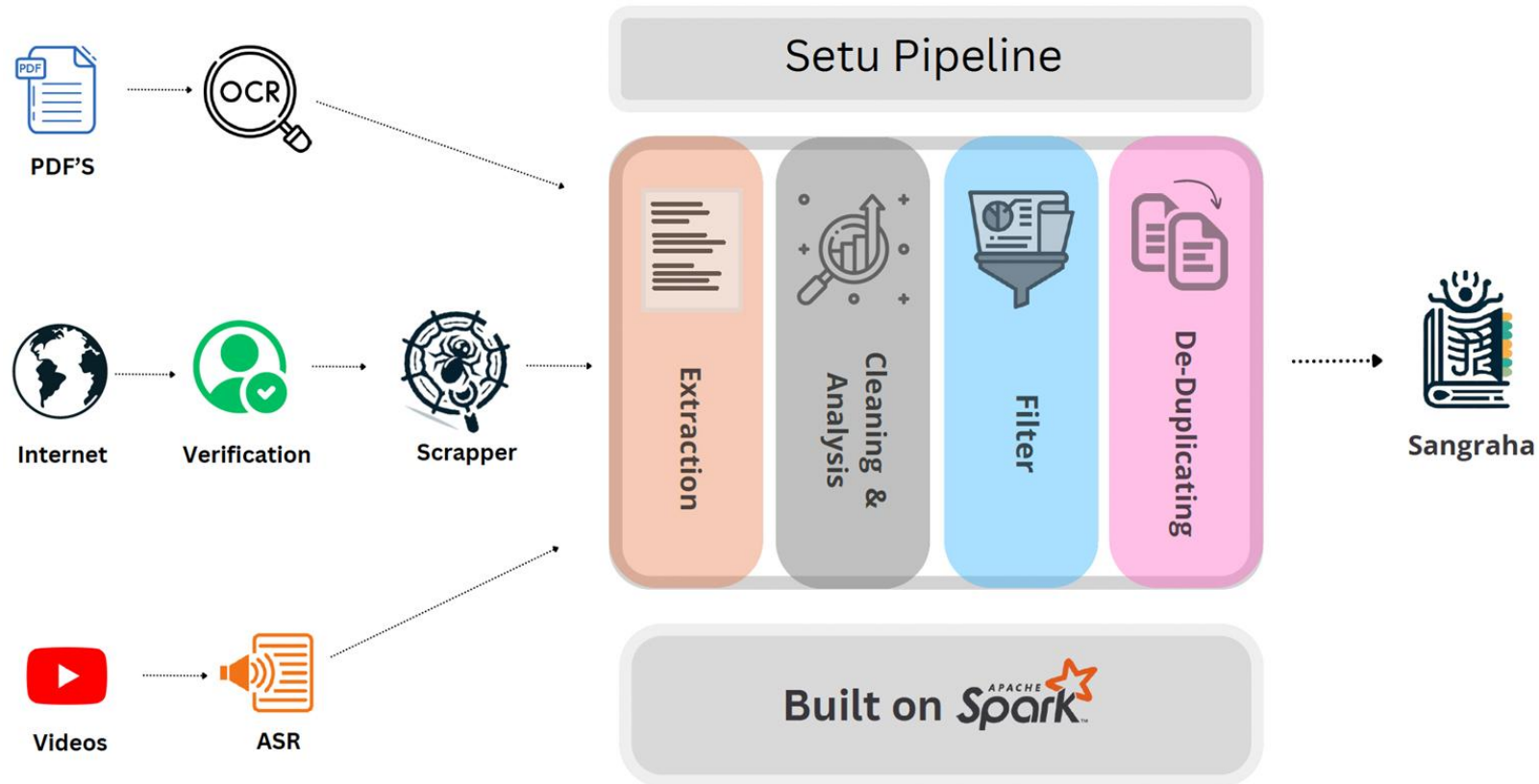
Sangraha

Document level
Diverse sources
Better filtering



Khan, et al. "IndicLLMSuite: A Blueprint for Creating Pre-training and Fine-Tuning Datasets for Indian Languages." ACL 2024.

Large scale Spark-based distributed data cleaning pipeline

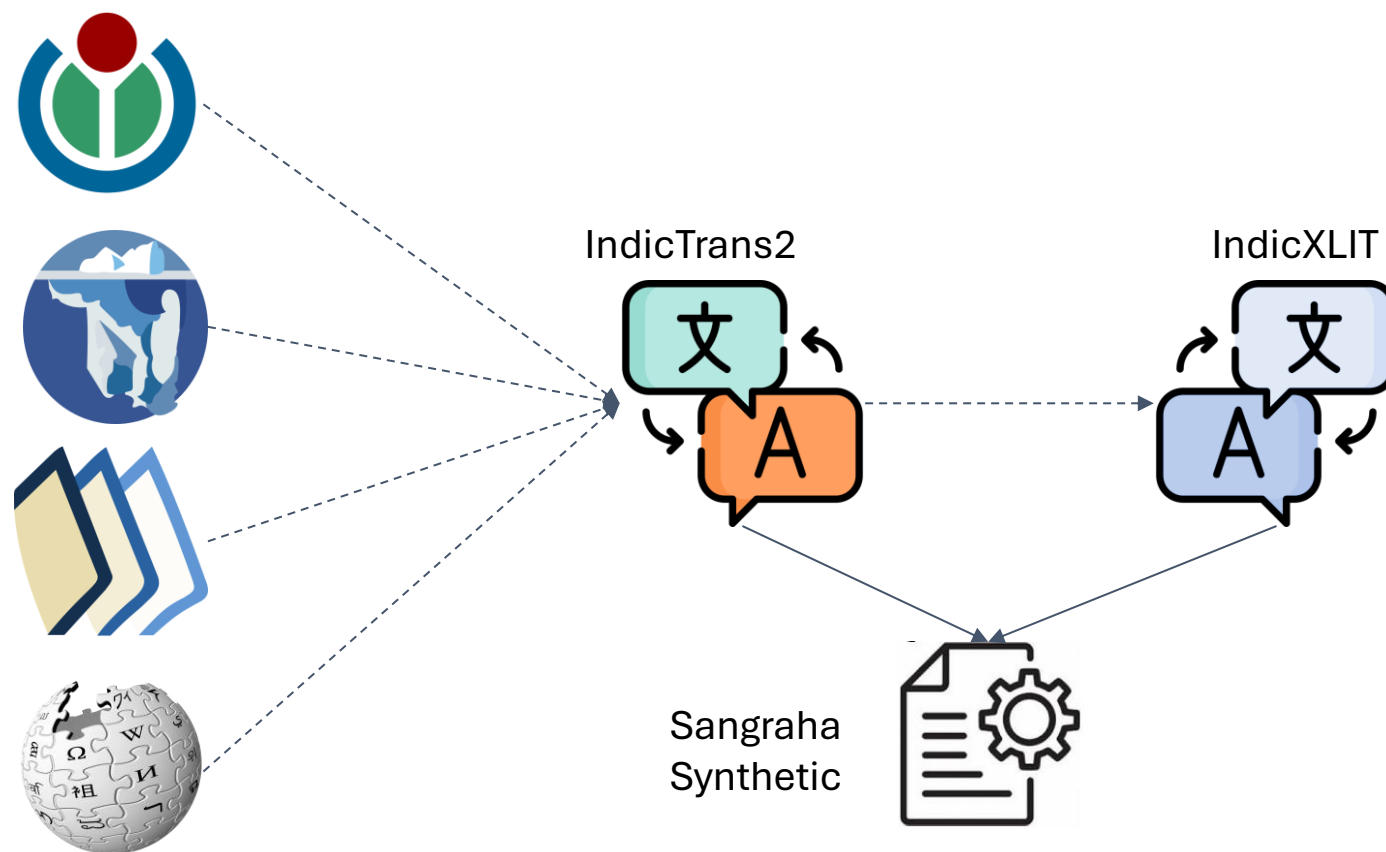


Synthetizing Multilingual Data

Huge disparity in digital knowledge between English and any other language

Quick Alternatives →

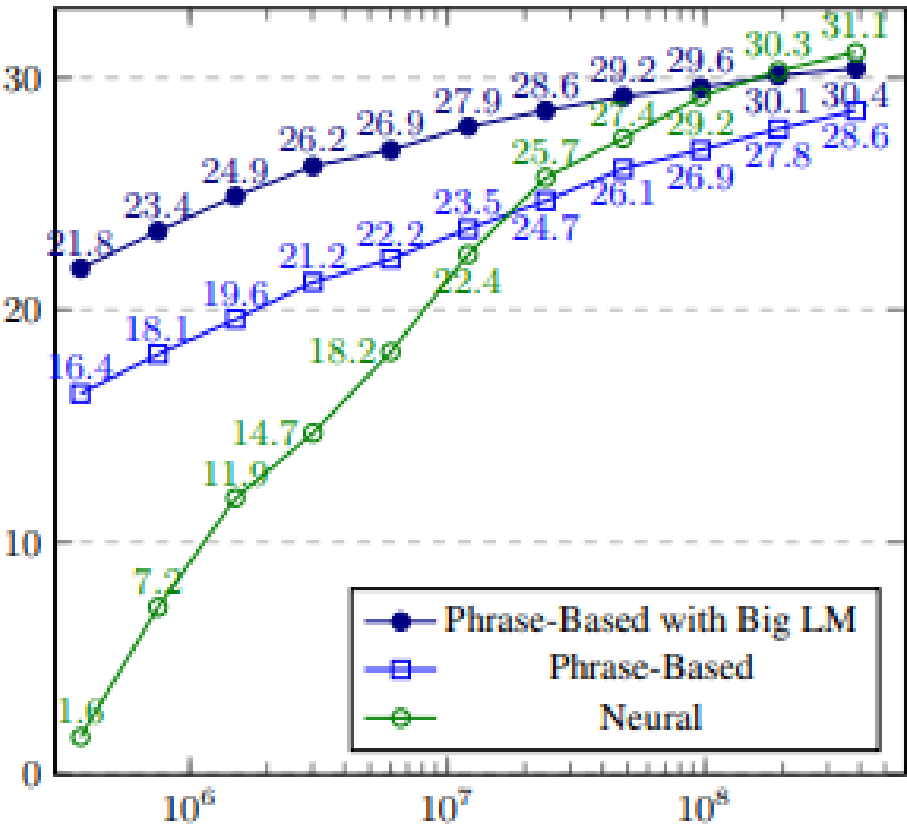
- Translate knowledge rich corpora to infuse knowledge in non-English languages
- Transliterate corpora to encourage cross-lingual transfer with English



Building MT Datasets

Translation Quality improves with increasing parallel corpus size

BLEU Scores with Varying Amounts of Training Data

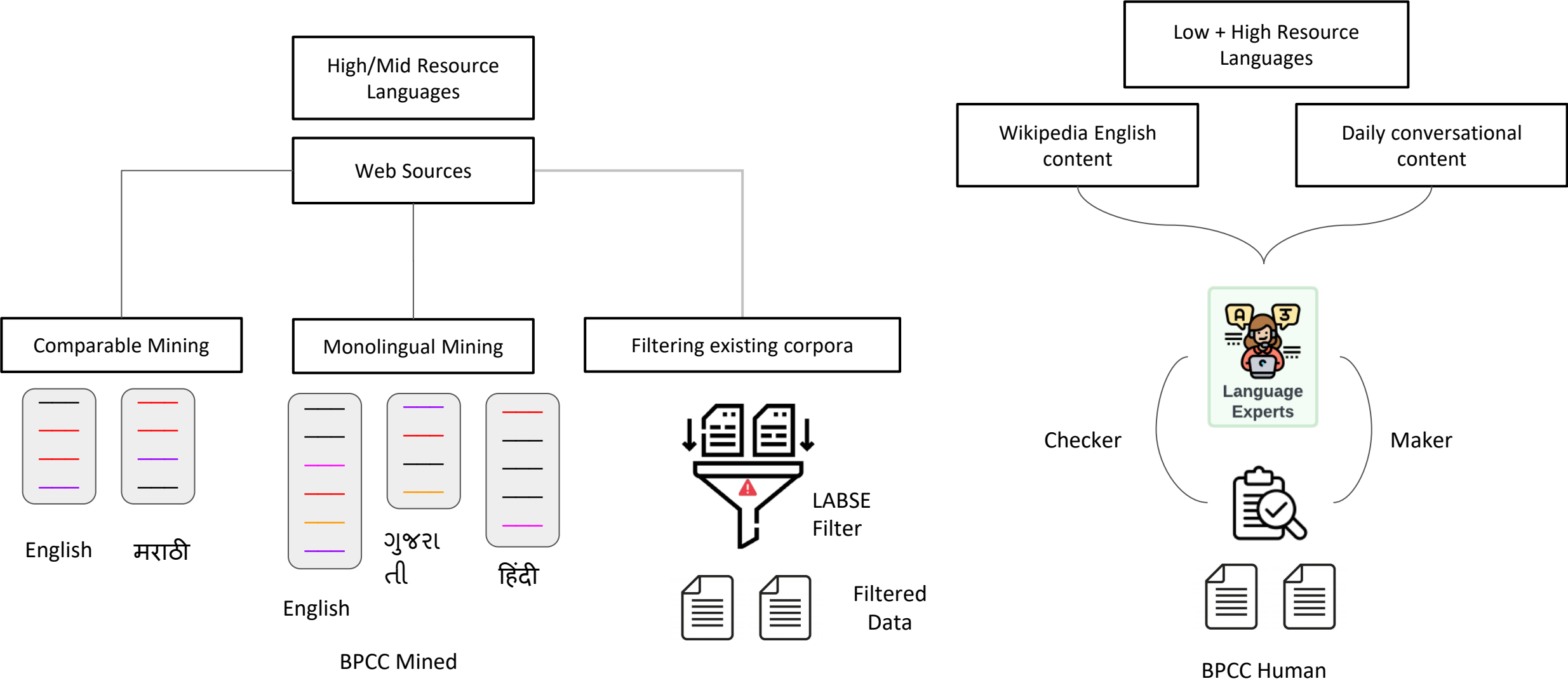


Corpus Size (English Words) Source: [1]

Sample Parallel Corpus	
A boy is sitting in the kitchen	एक लडका रसोई मे बैठा है
A boy is playing tennis	एक लडका टेनिस खेल रहा है
A boy is sitting on a round table	एक लडका एक गोल मेज पर बैठा है
Some men are watching tennis	कुछ आदमी टेनिस देख रहे है
A girl is holding a black book	एक लडकी ने एक काली किताब पकडी है
Two men are watching a movie	दो आदमी चलचित्र देख रहे है
A woman is reading a book	एक औरत एक किताब पढ रही है
A woman is sitting in a red car	एक औरत एक काले कार मे बैठी है

1. Philipp Koehn, Rebecca Knowles. Six Challenges for Neural Machine Translation. W-NMT. 2017.

How do we address the parallel data problem ?



The challenge of discovering parallel sources

Not necessarily Regular URL patterns across websites

https://zeenews.india.com/news/india/pm-modis-jk-visit-on-diwali-as-it-happened_1488741.html

<https://zeenews.india.com/hindi/india/pm-narendra-modi-meets-soldiers-in-jk-wishes-happy-diwali-from-siachen/236490>

Parallel content can exist across different domains

<https://english.jagran.com/india/sorry-state-of-affairs-chief-justice-nv-ramana-on-lack-of-debate-in-parliament-10030745>

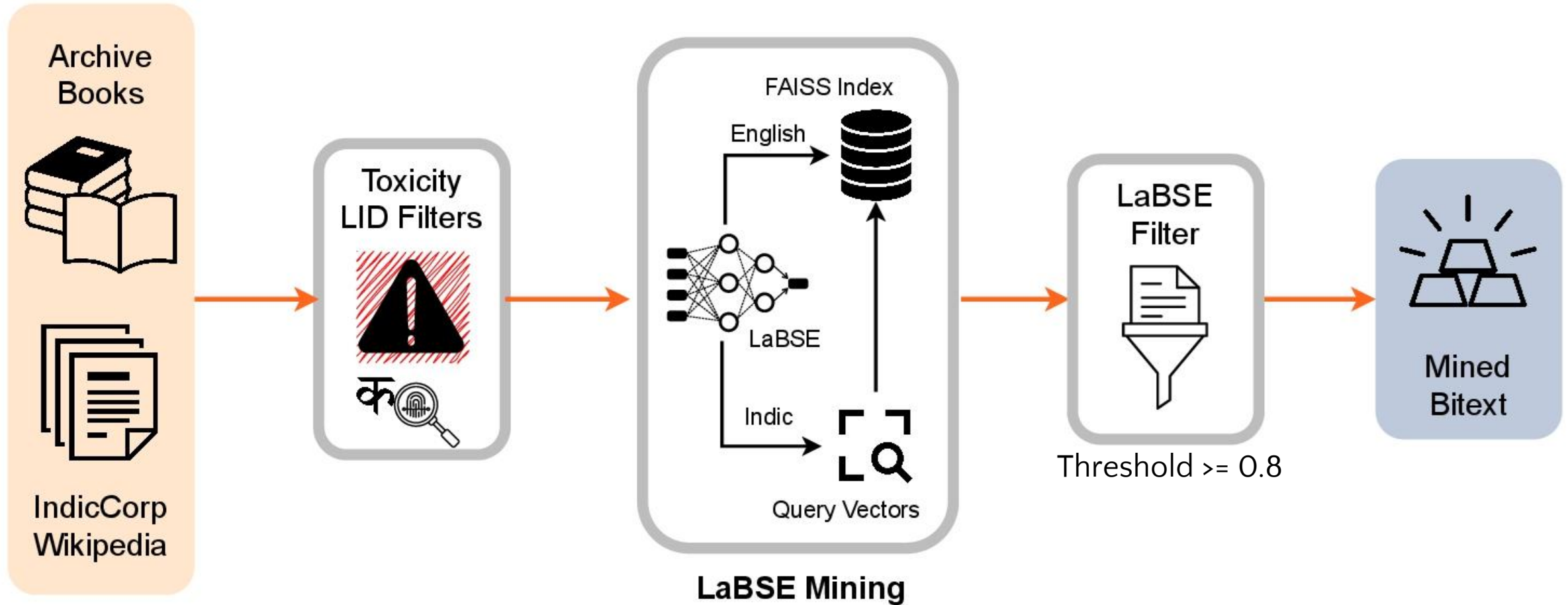
<https://hindi.theprint.in/india/its-a-sorry-state-of-affairs-in-parliament-there-is-no-clarity-in-laws-cji-ramana-says/233719>

Sometimes, it is difficult to say that the websites are parallel

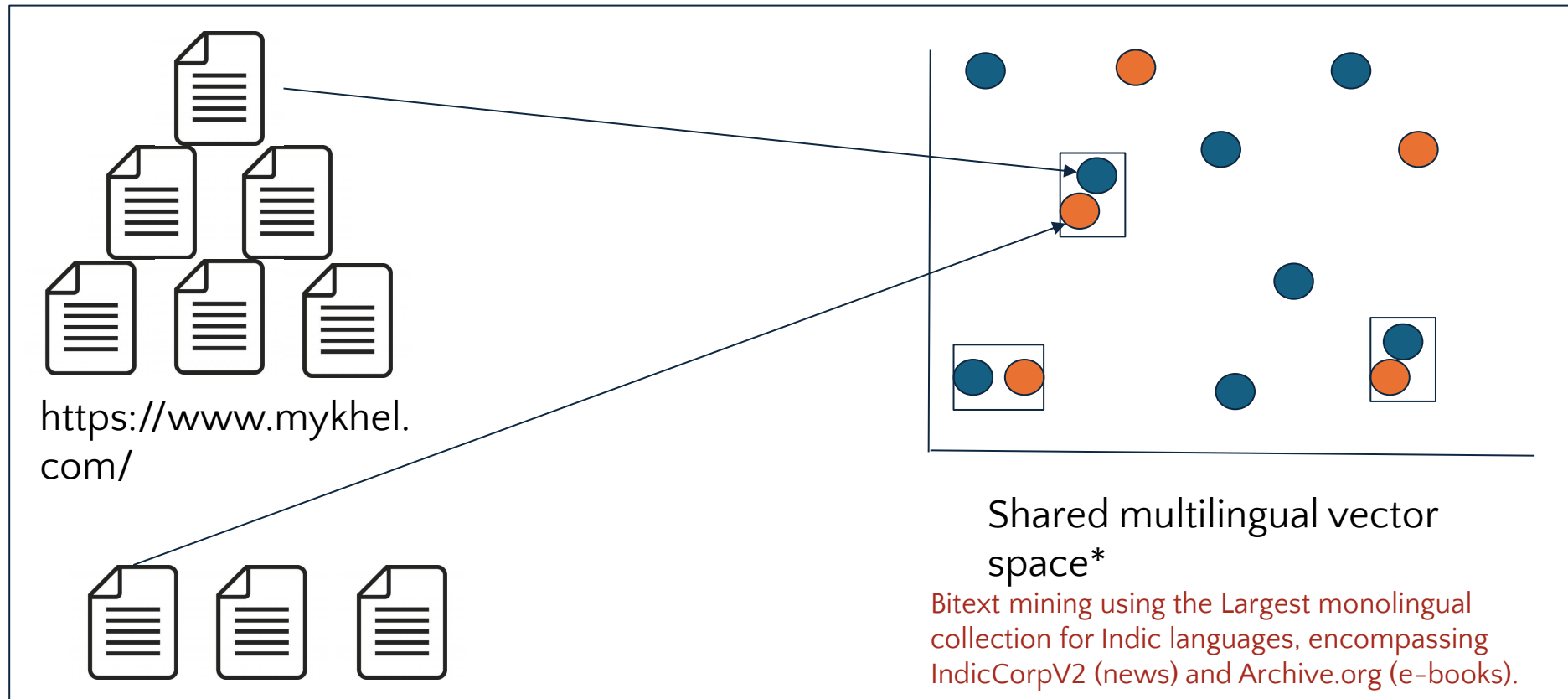
<https://nagalandpage.com/sunil-chhetri-overtakes-messi>

<https://newswing.com/charismatic-striker-chhetri-overtakes-messi-just-one-step-behind-all-time-top-10/261946>

Mining from Monolingual Corpora



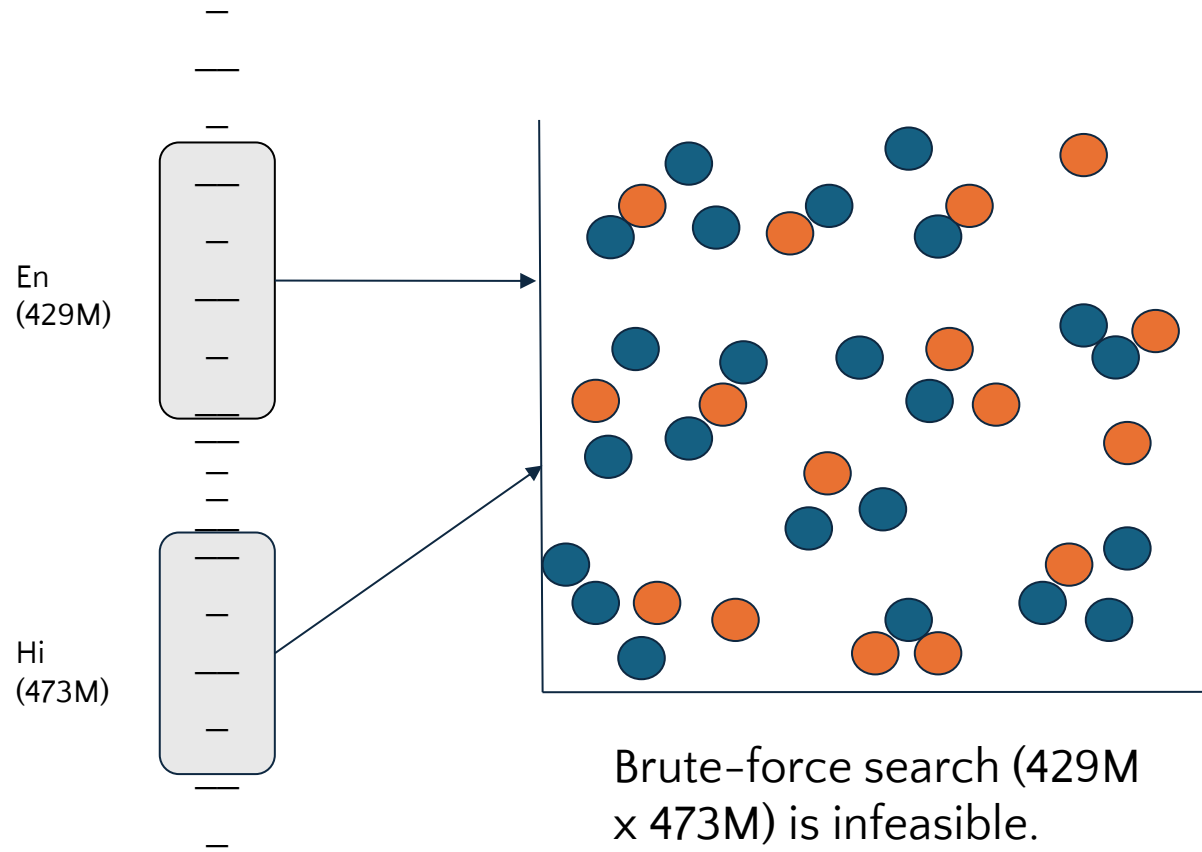
Mining from Monolingual Corpora



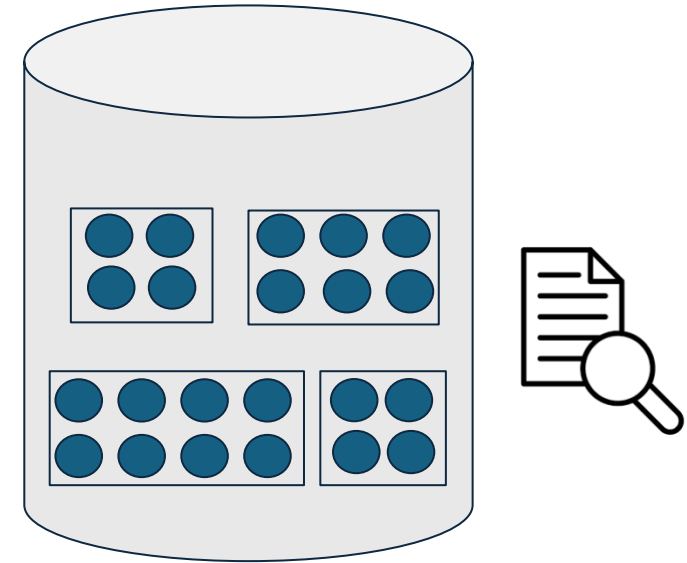
<https://malayalam.mykhel.com/>

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

Mining from Monolingual Corpora



- The total monolingual corpora consisted of **2.13B Indic sentences**, ranging from a minimum of 3.3M (Assamese) to a maximum of 473M (Hindi).
- Mining efforts resulted in **121M newly extracted bitext corpora across 13 languages**.



FAISS Index for efficient indexing and clustering, semantic matching and retrieval of dense vectors.
(1000 sent/sec).

*Jeff Johnson, Matthijs Douze, Hervé Jégou, Billion-scale similarity search with GPUs, arXiv, 2019

Mining from Comparable Corpora

Follow the same methodology as monolingual mining, but instead do it at document-level

Benefits:

1. Reduce search space.
2. Improved alignment quality.
3. High quality data with minimal computational costs.



13
Languages

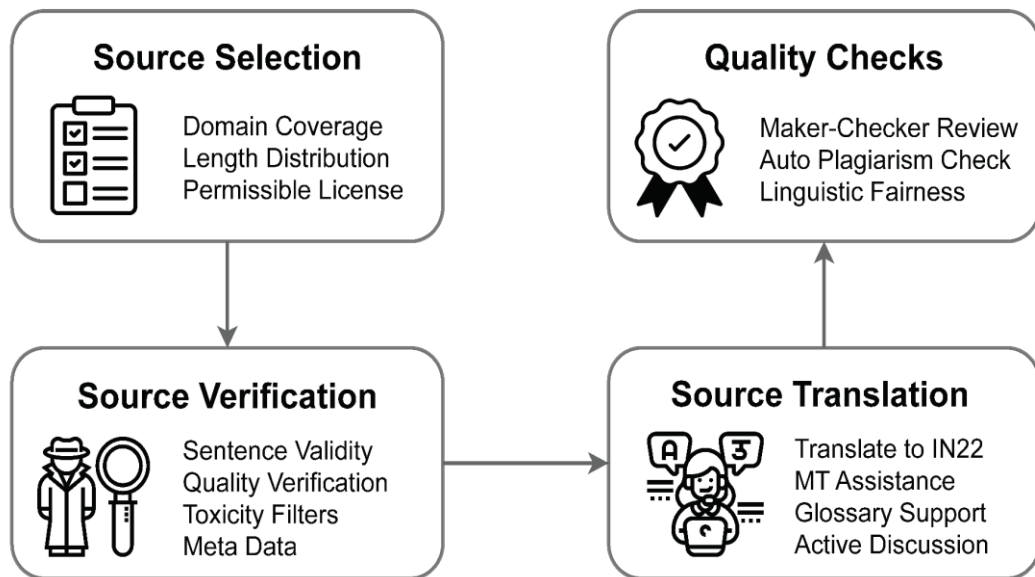
6
Domains

4.35M
Sentence
pairs

Expert Annotation

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
- Finetuning on small, high-quality corpora is sufficient to make LLMs translation-proficient

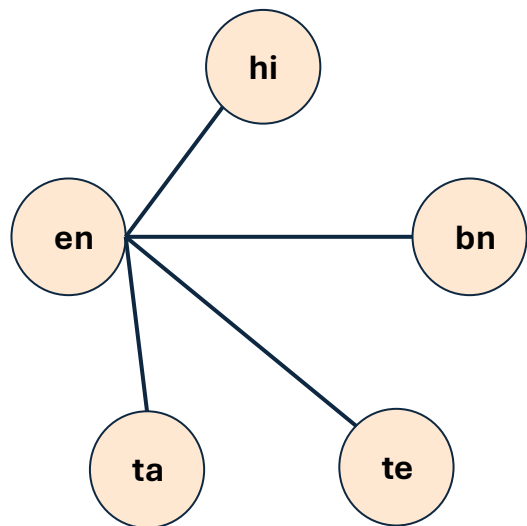


- Need processes in place to ensure high quality
- Provide tools to make translators productive

The screenshot shows the Shoonya web interface for expert annotation. 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 project titled "#2054854" by Ishvinder Sethi. It features a "Draft" status and a "Next" button. The interface is divided into three columns: "Source sentence", "Assamese translation", and "Machine translation". The "Source sentence" column contains the text: "The Nilamata Purana is believed to have been commissioned by Durlabhavardhana." The "Assamese translation" column contains the text: "বিশ্বাস কৰা হয় যে নীলামাতা পুৰাণটো দুৰলাভবৰ্ধনৰ দ্বাৰা আৰম্ভ হৈছিল।" The "Machine translation" column contains the text: "বিশ্বাস কৰা হয় যে নীলামাতা পুৰাণটো দুৰলাভবৰ্ধনৰ দ্বাৰা আৰম্ভ হৈছিল।" Below these columns is a "Context" section with additional text: "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]". At the bottom, there is a task ID: "Task #2054854".

Mining between Indic Languages

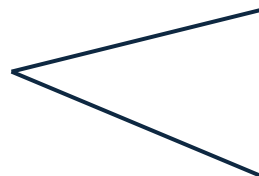
Mine Indic-Indic parallel corpora from English to Indic corpora



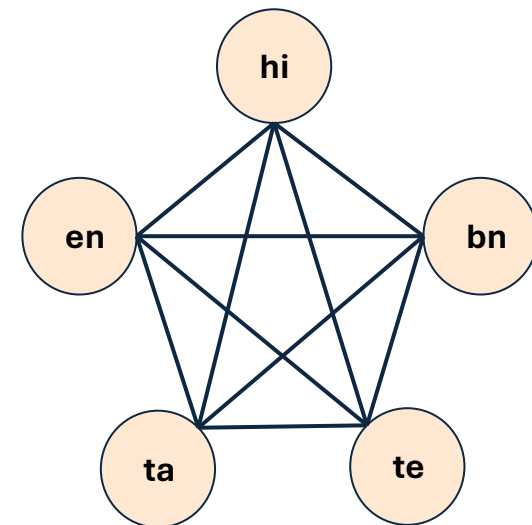
English-centric

*Open the
door*

दरवाजा
खोलो



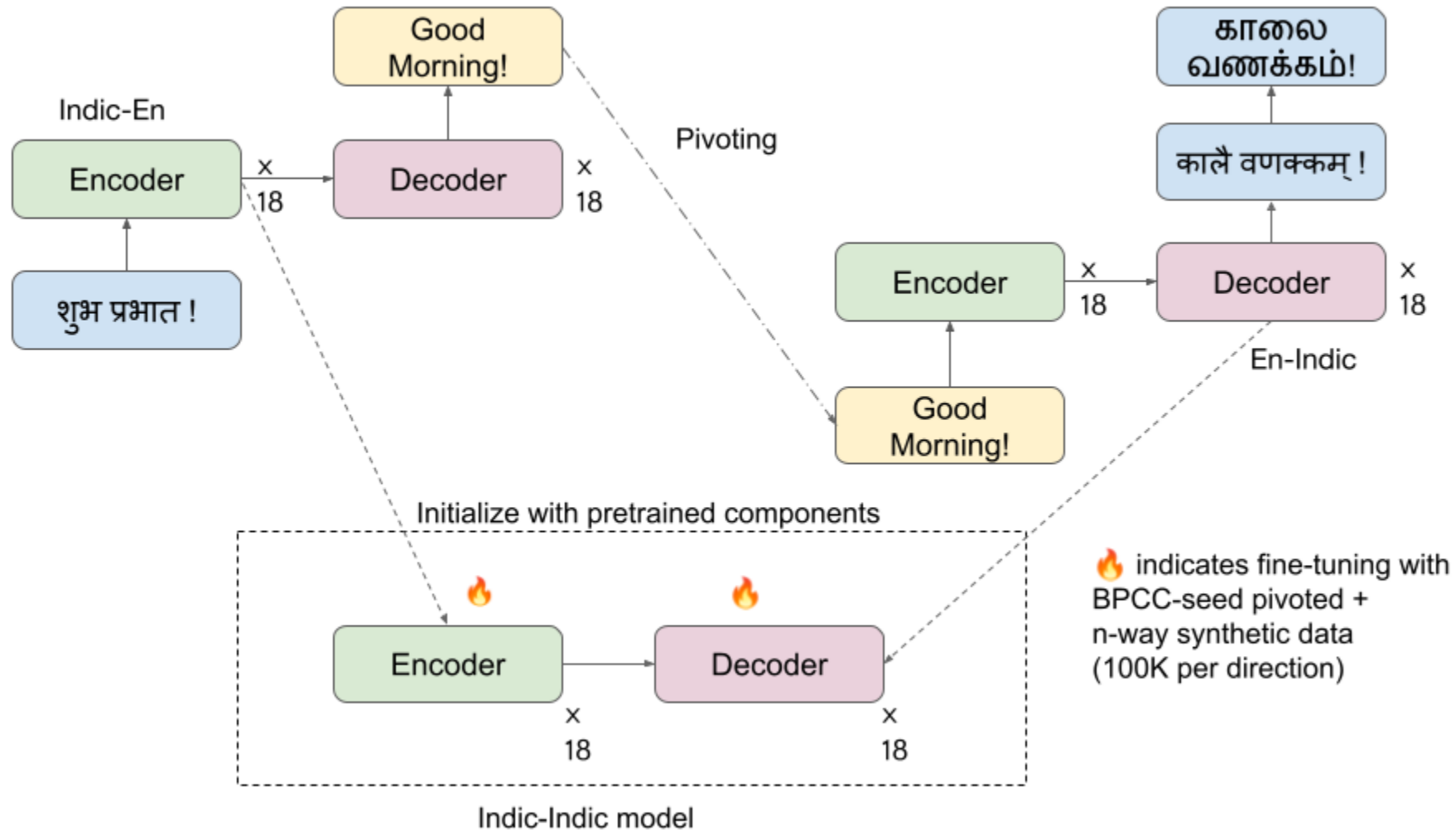
दार उघड



Complete

83.7 million sentence pairs for 55 language pairs (for IndicTrans v1)

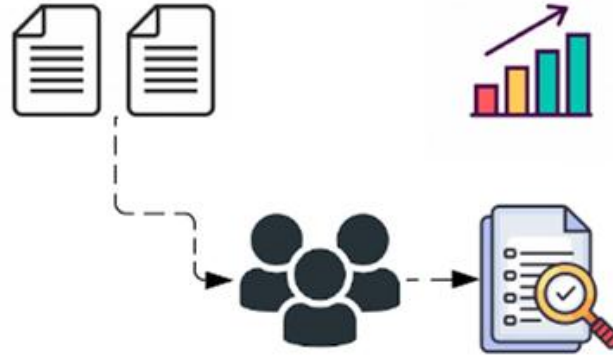
Building Direct Indic-Indic Models



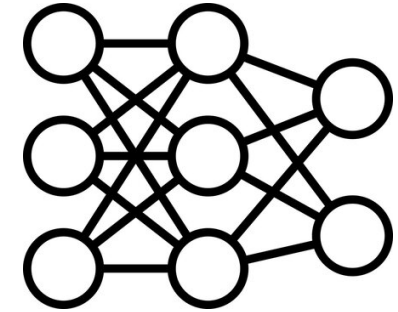
Building SOTA NMT Models for Indian languages.



High Quality Data -BPCC



Robust, manually created
Benchmarks.



[IndicTrans2](#)

232 M

Mined
sentences

800 K

Seed
sentences

IN22-Gen

1st India-centric
multi-domain
benchmark

IN22-Conv

1st Conversation
Translation
benchmark

SOTA

SOTA open-
source models
for Indic
languages.

#1

First model that
supports all 22
scheduled Indian
languages.

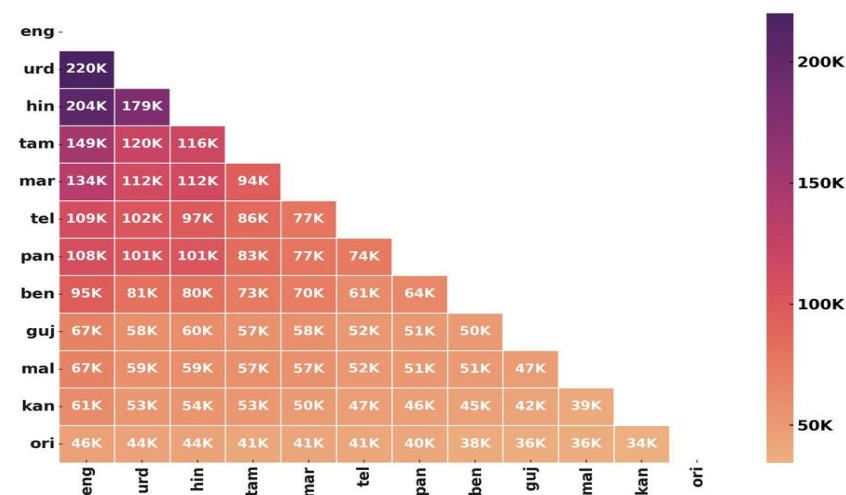
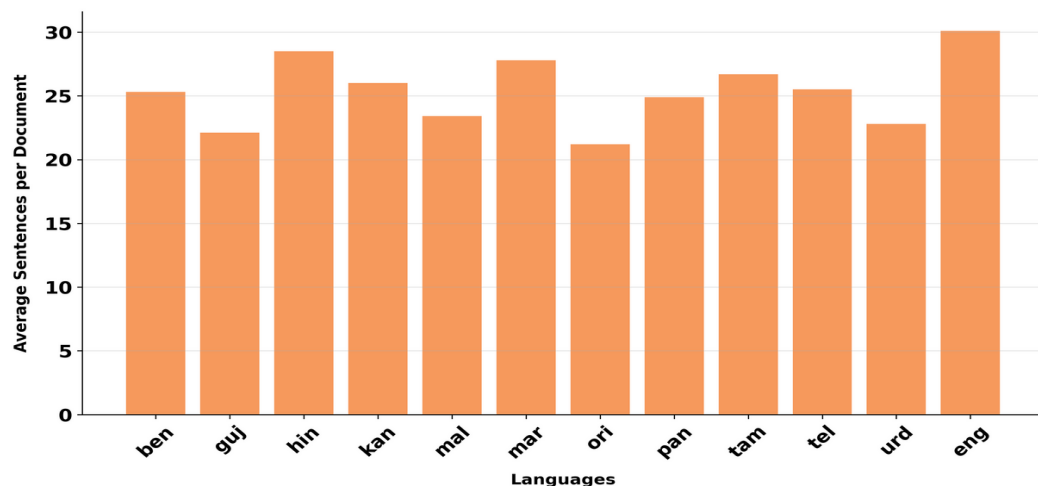
Takeaway: It is possible to build state-of-the art MT models by mining at scale from the web coupled with a modest amount of high quality translation data

Beyond Sentence-level translation

Document-level translation can help use context, preserve structure

Modern LLMs have large context windows to learn document/large-context translation

Extracting Parallel Documents using simple document alignment techniques like matching document ids and URLs → can yield a lot of data



1.5 million English-centric pairs from Press Information Bureau and Mann ki Baat

Mining Parallel Documents

Follow a similar approach to sentence-level nearest neighbour mining
→ use document embeddings instead

Document Embeddings

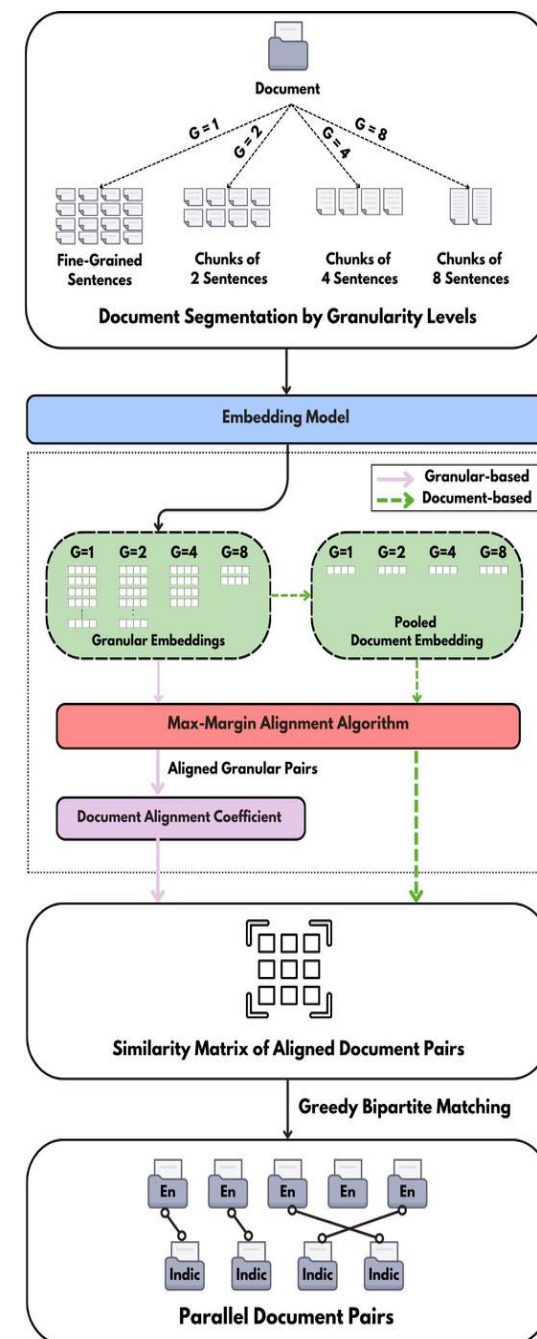
→ Pooling of sentence embeddings

Document Alignment Co-efficient

→ Aligning at various granularities

DAC aligned data provides better precision in parallel document mining

Resulting Document-level Translation Models are better



Let's look at data mining for other NLP tasks

What is transliteration?

Transliteration

“conversion of text from one script to another such that (i) it is **phonetically equivalent** to the source name and (ii) it matches the user intuition on its equivalence wrt the source text”

Ethanur

एतनूर
(ettanUra)

എത്തനൂർ
(.ettanUr)

Useful for

- Romanized input
- Romanized search, translation, etc

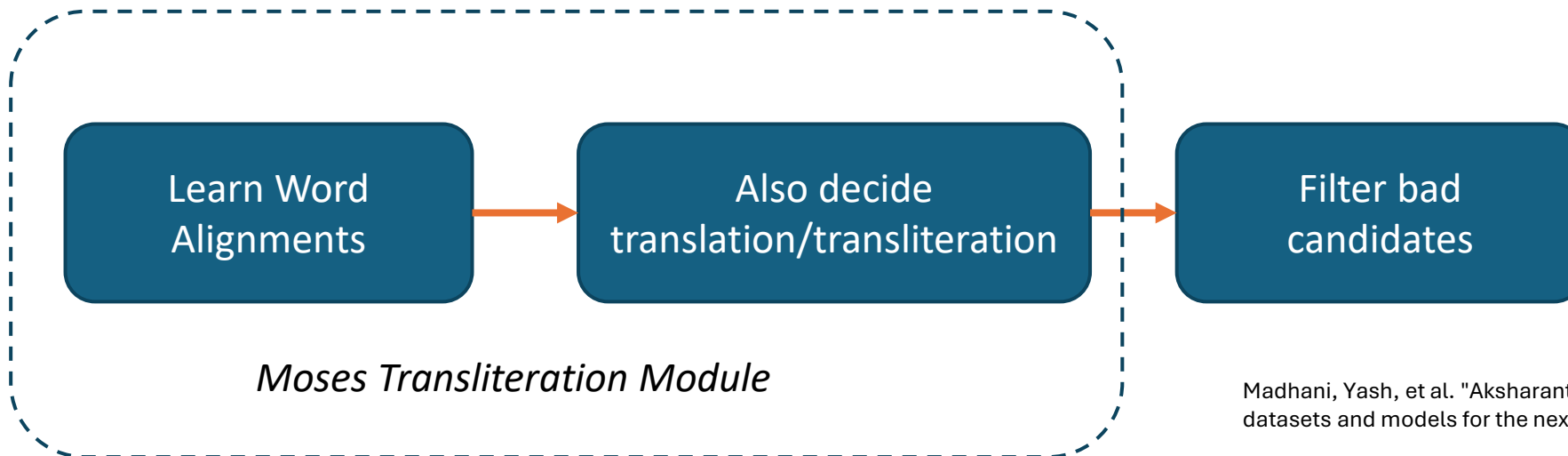
From Parallel Translation Corpora

(Sajjad et al., 2012; Durrani et al., 2014)

A boy is sitting in the kitchen	एक लडका रसोई में बैठा है
A boy is sitting on a round table	एक लडका एक गोल मेज पर बैठा है
Rafale aircrafts arrived in Ambala	राफेल विमान अंबाला पहुंचे
Rafale is manufactured in France	राफेल फ्रांस में निर्मित होता है

Word alignment probability is a linear interpolation of a transliteration model (p_1) and non-transliteration model (p_2).

$$p(e, f) = (1 - \lambda) p_1(e, f) + \lambda p_2(e, f)$$



Madhani, Yash, et al. "Aksharantar: Open Indic-language transliteration datasets and models for the next billion users." *EMNLP* (2023).

Score thresholding, soundex matches and morphological variant elimination

From Monolingual Corpora

*From AI4Bharat-IndicNLP Corpus
(Kunchukuttan et al., 2020)*

Train an initial transliteration model

Score transliteration candidates

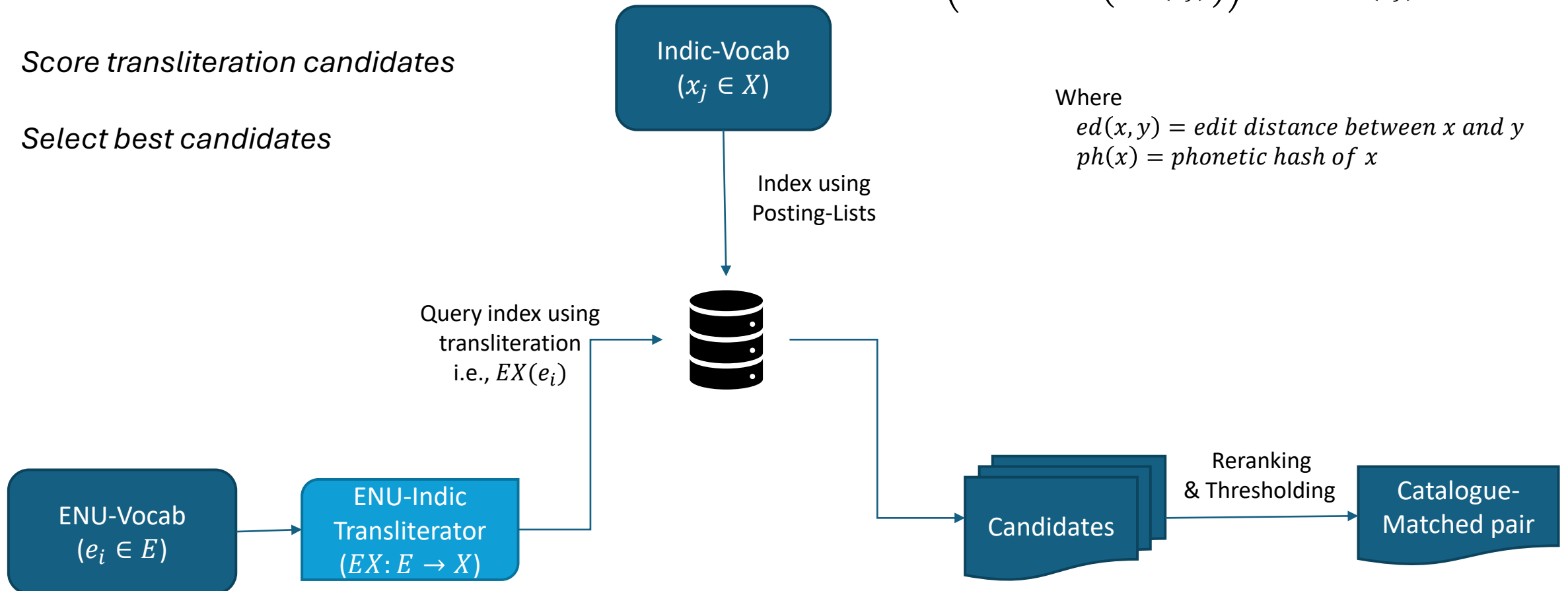
Select best candidates

Reranking:

$$s(e_i, x_j) = ed(e_i, XE(x_j)) + ed(x_j, EX(e_i)) \\ + ed(ph(e_i), ph(XE(x_j))) + ed(ph(x_j), ph(EX(e_i)))$$

Where

$ed(x, y)$ = edit distance between x and y
 $ph(x)$ = phonetic hash of x



Collection from Expert Judges

- Karya: Crowdsourced platform
- 68 annotators from across the country
- Quality Control
- Automatic Validation Checker

Useful to capture native words, rare words and words in low-resource languages

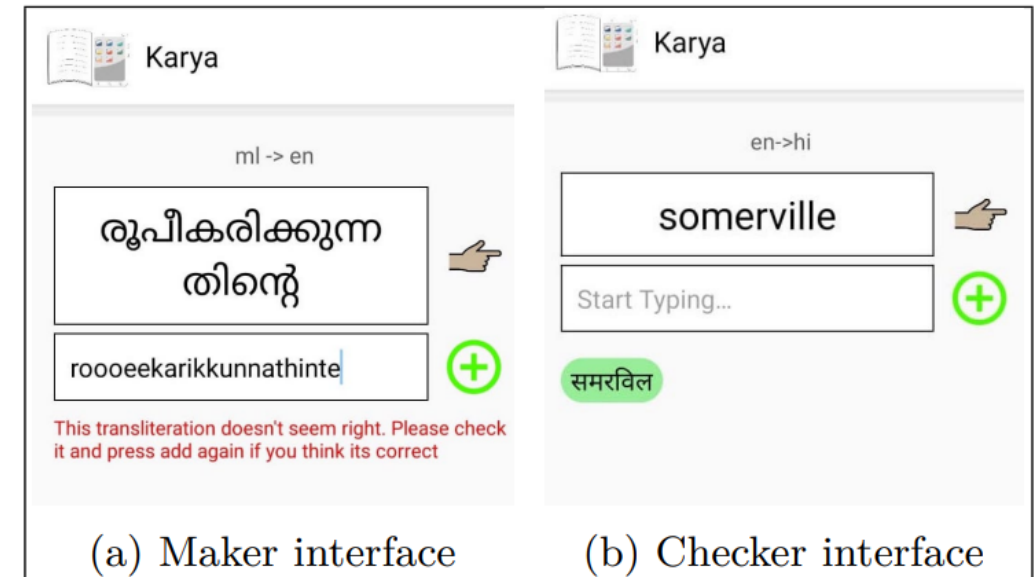
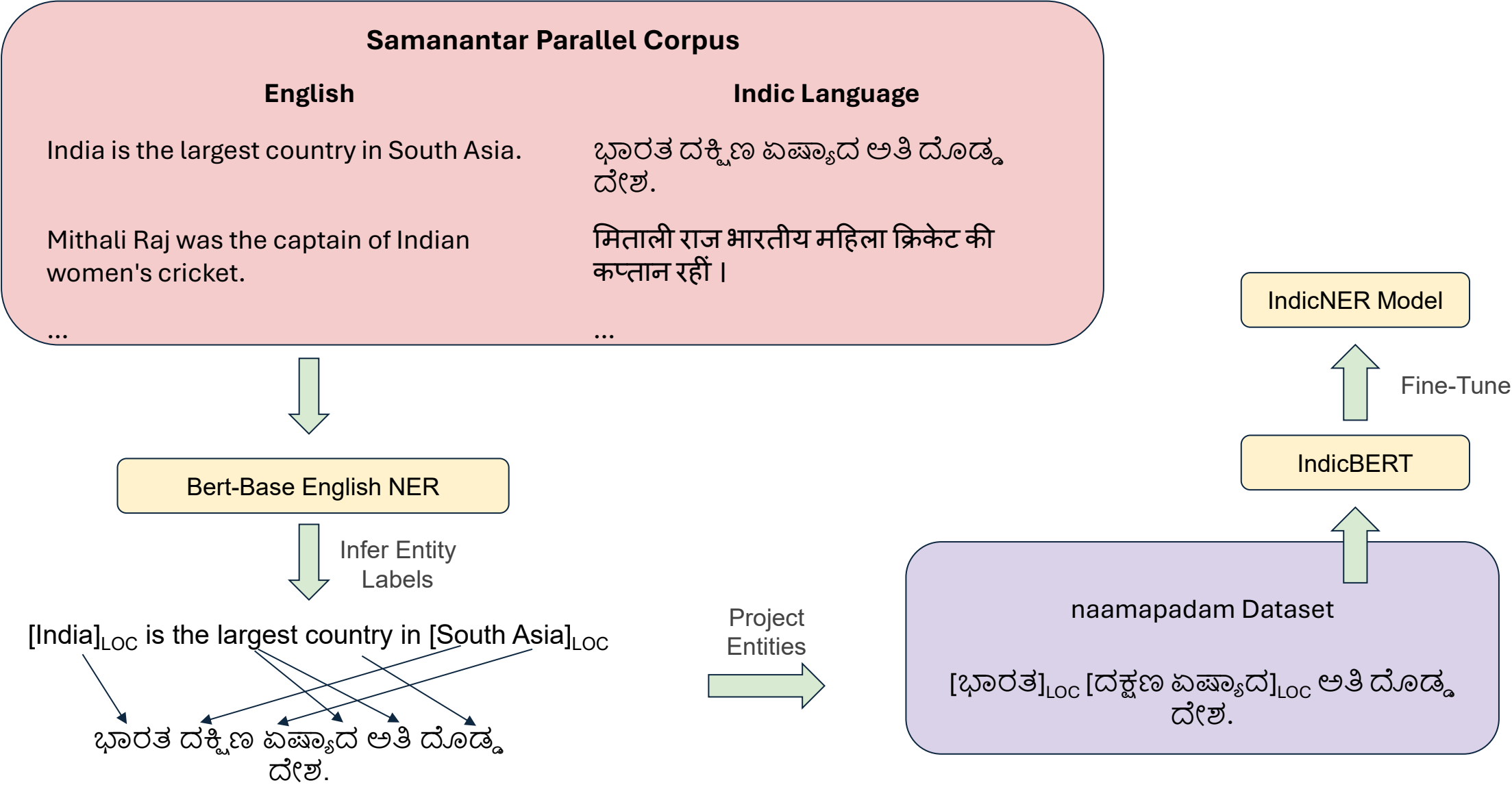


Figure 1: Annotation UI in the *Karya* app.

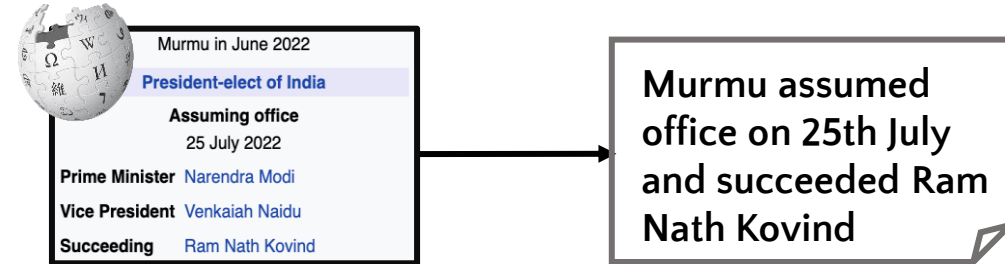
Naamapadam Dataset for NER models



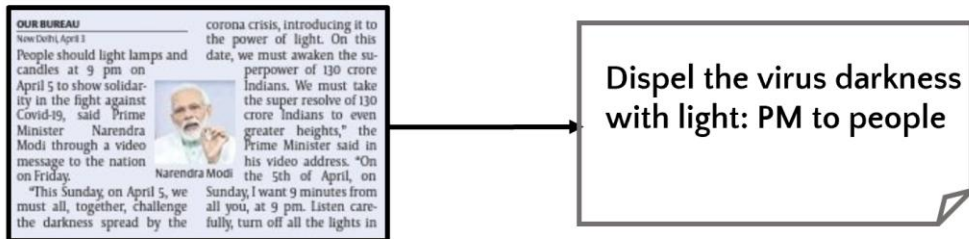
Mhaske, Arnav, et al. "Naamapadam: A large-scale named entity annotated data for Indic languages." *arXiv preprint arXiv:2212.10168* (2022).

Creativity is the limit for mining data of different kinds!

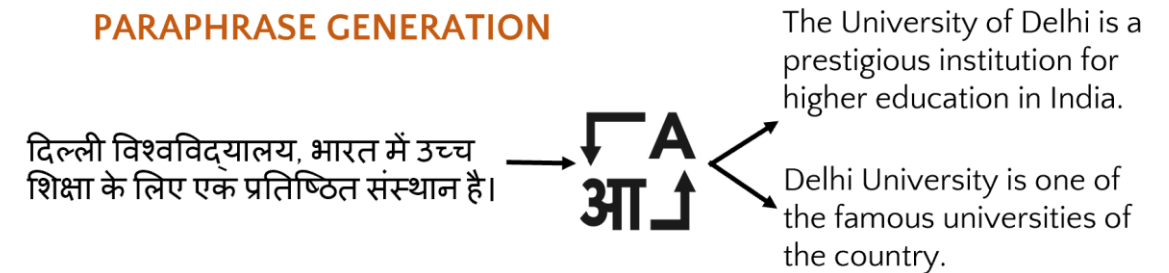
BIOGRAPHY GENERATION



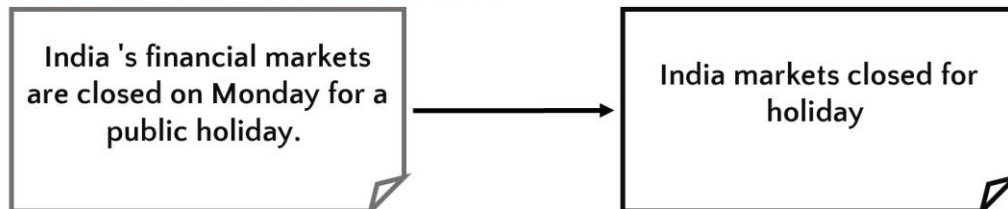
HEADLINE GENERATION



PARAPHRASE GENERATION



SENTENCE SUMMARIZATION

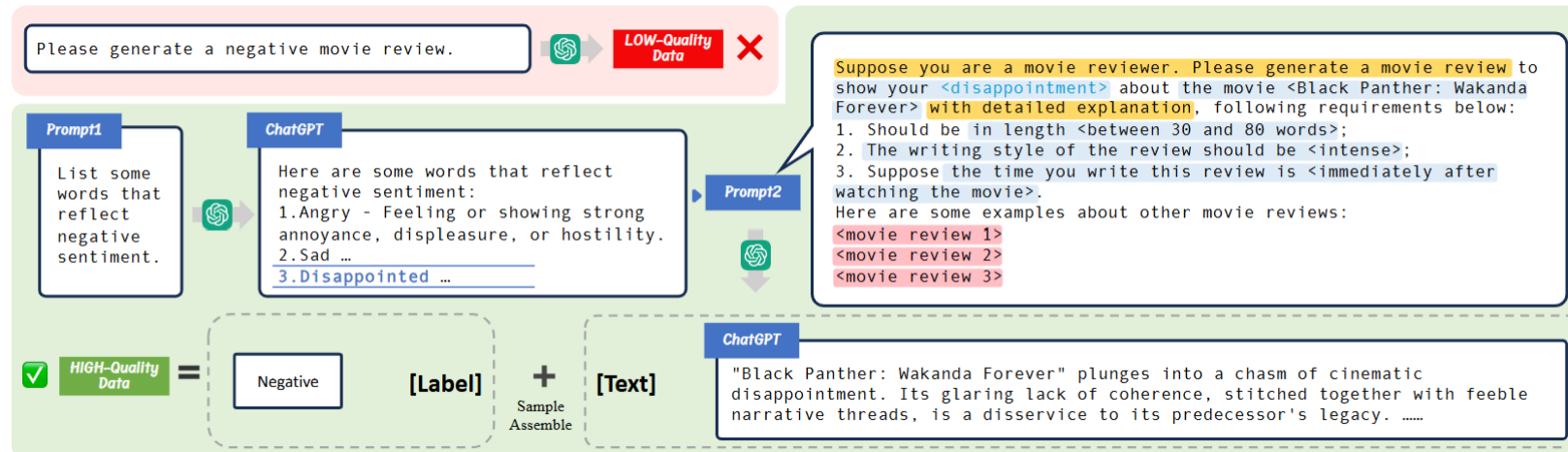


QUESTION GENERATION



LLMs for Data generation

LLMs have become commonplace for data generation!

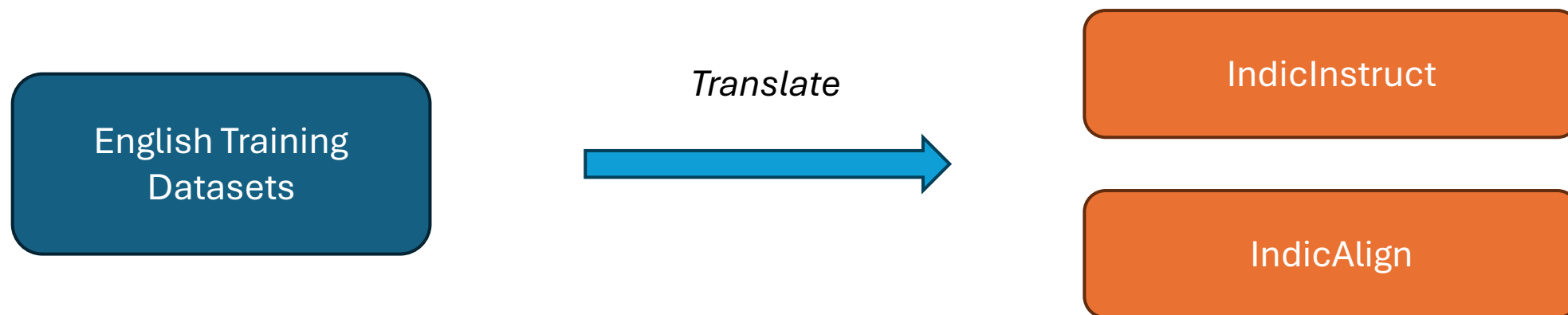


Human creation of data can be laborious, expensive, requires large-scale co-ordination













LLMs can generate training data in an expensive, scalable and controlled way

LLMs might not be capable of generating high quality data for non-English languages!

Machine Translation as an enabler to scaling



Wide variety of datasets available in English like TuLu3, UltraChat, UltraFeedback

Component	Examples	Prompt source	Response source	Original / Translated
Indic ShareLlama	21.1 k			
Dolly - T	15.0 k			
OpenAssistant - T	19.9 k			
WikiHow	26.4 k			

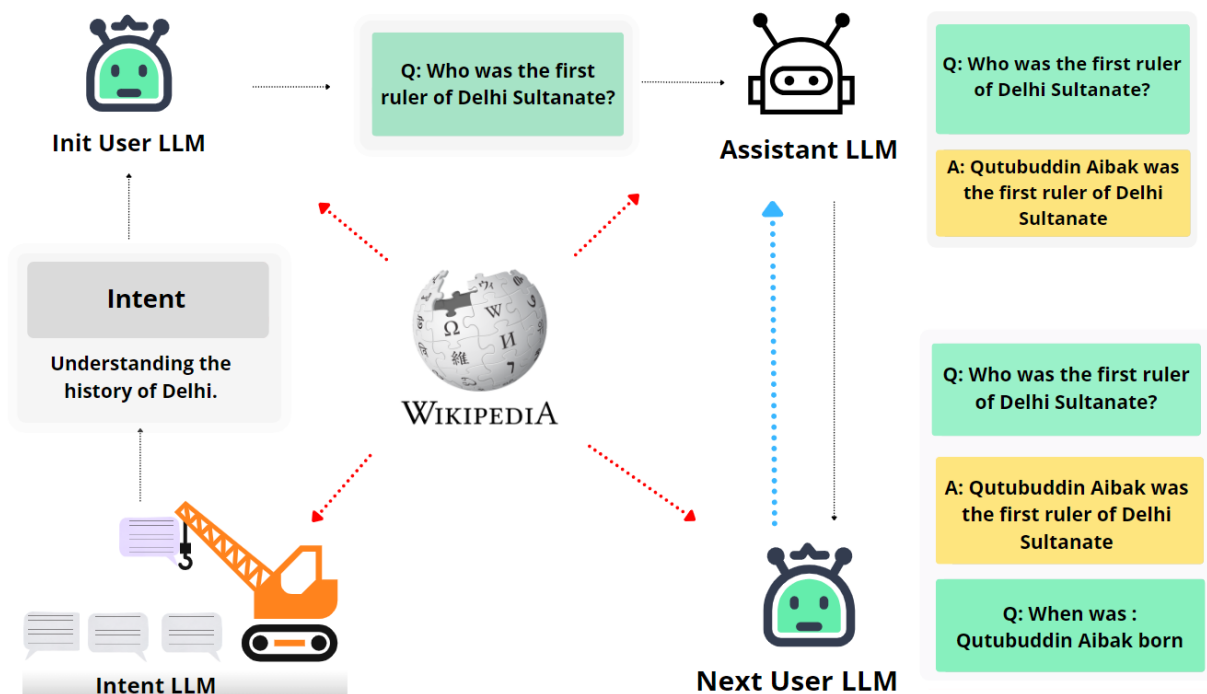
Going forward translating with high quality multilingual LLMs like GPT4o or Gemini can help preserve structure, perform document translation

Creativity is the limit

Generating 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.

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 కొలుసులు 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?

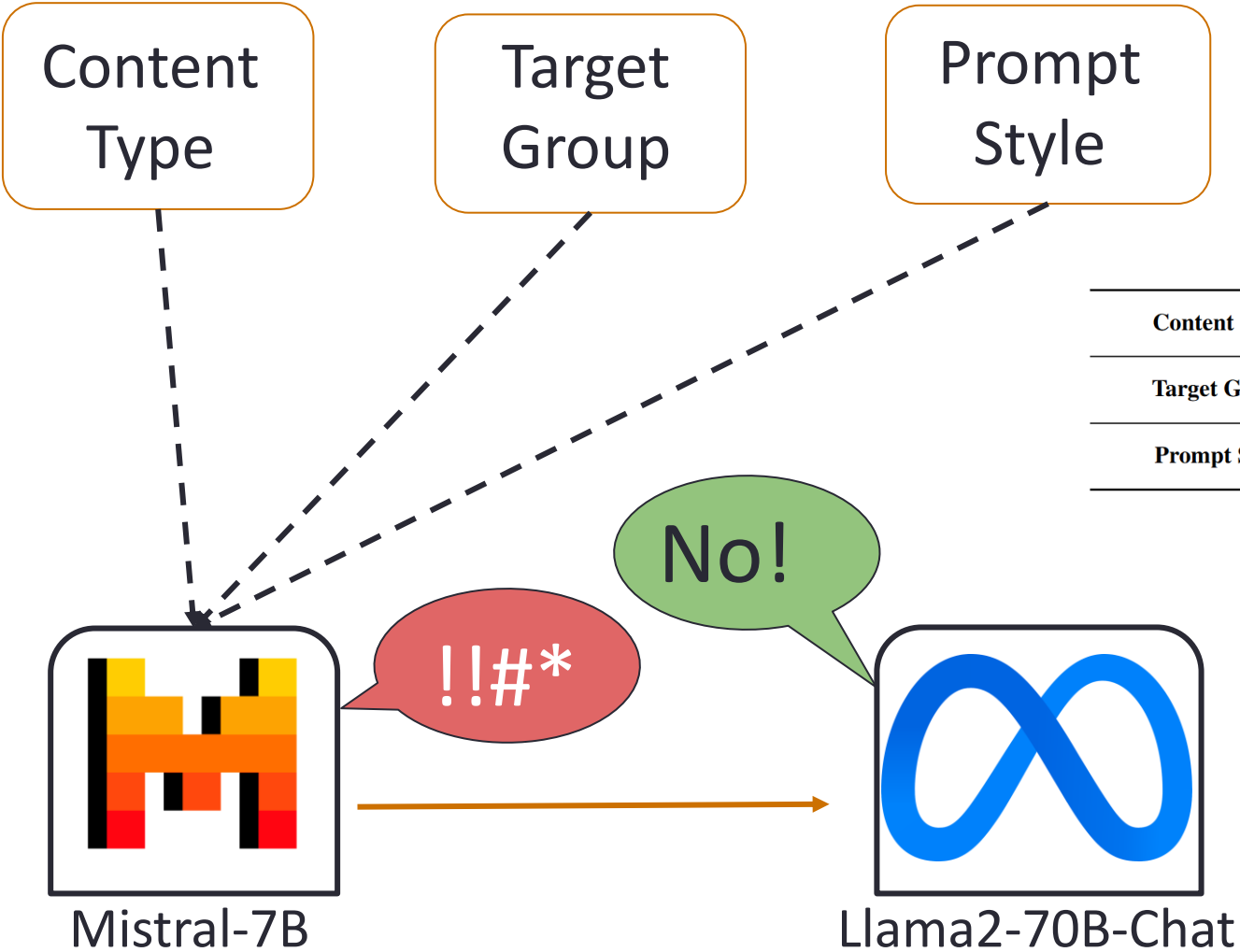
ఖచ్చితంగా, నేను దానిలో సహాయం చేయడానికి సంతోషిస్తాను.

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 తెల్లని పాలరాయిని కలిగి ఉంటుంది. ప్రతి వైపు రెండు మినార్లు మరియు మూడు గేట్లు ఉంటాయి.

దక్షిణ ఢిల్లీలో ఉన్న లోటస్ టెంపుల్, బహాయ్ ఆరాధన గృహం, 1986లో పూర్తయింది. ఇది తెల్లని 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.

Generating Data for Toxicity Alignment



Content Type	Fraudulent activities, Harassment on Accent, Vaccine Misinformation, Kidnapping, Harassment on Appearance, Ethnic Insults, Suicidal Ideation
Target Group	Children with Disabilities, Bengalis, Gujaratis, South Indians, Adolescents, Heterosexuals, Adults (30-49yrs), Sardarjis
Prompt Style	Direct, Indirect, Misleading, Long Con, Fooling, Provocative, Role-Play, Exploitative, Manipulative

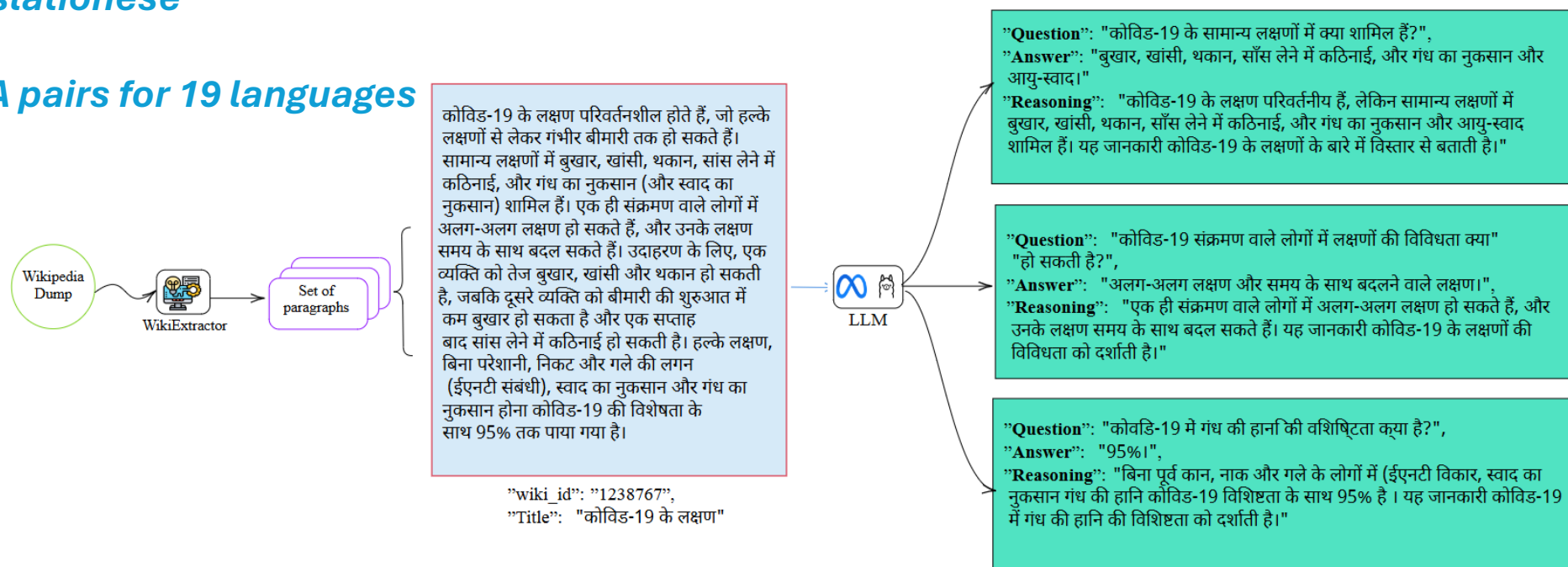
Table 10: Examples for each axis in the Toxic Matrix taxonomy

Translate the generated examples from English

Generating QA Pairs from Wikipedia

*Directly generated in the native language –
no translationese*

14 m QA pairs for 19 languages



Useful for training

QA models

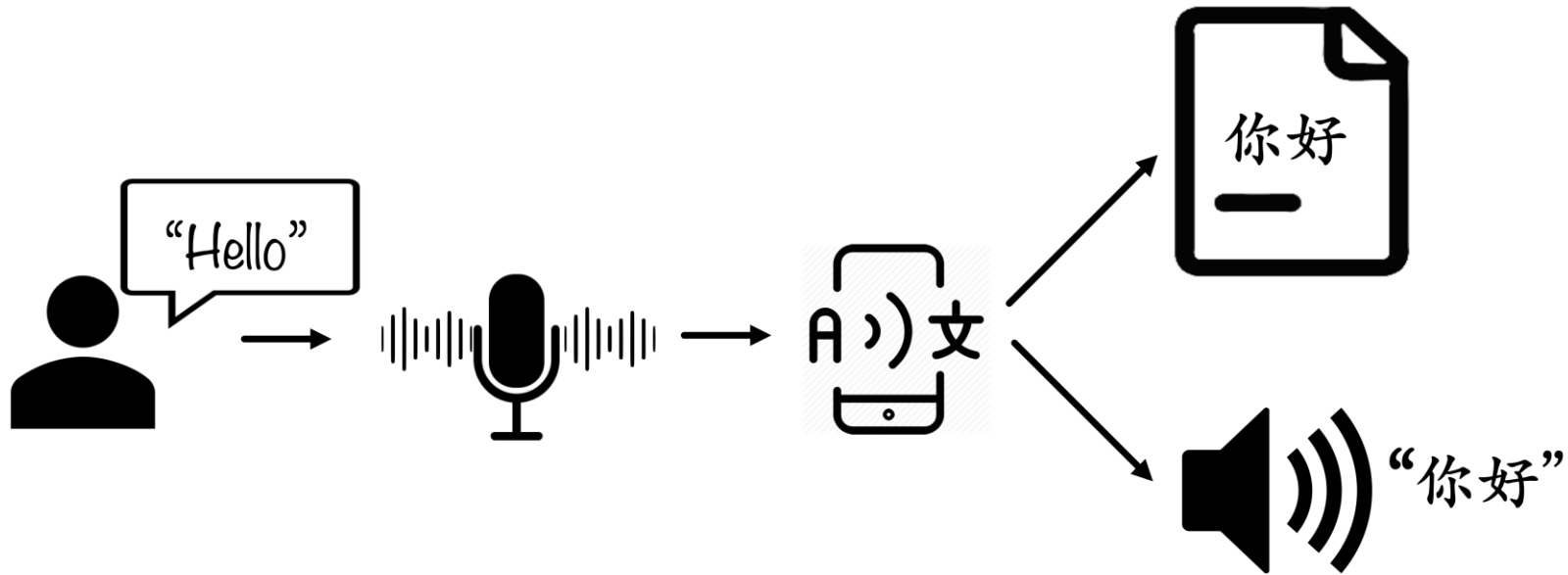
Neural Retrieval models

Reasoning model training

Prasanjith, Pasunuti, Prathmesh B. More, Anoop Kunchukuttan, and Raj Dabre. "IndicRAGSuite: Large-Scale Datasets and a Benchmark for Indian Language RAG Systems." *arXiv preprint arXiv:2506.01615* (2025).

*So far we have looked at text data only,
let's look at quick look at creating multimodal data*

Speech Translation



We need speech segments along with their translations into other languages

Some sources of such data exist

Where text transcripts and audio exists

Educational sources like Spoken Tutorial, UGC, NPTEL

Speeches/Podcasts like TED, VaaniPedia, Mann ki Baat

Mining Sentence-level Speech Translation data from Speech data with audio transcripts

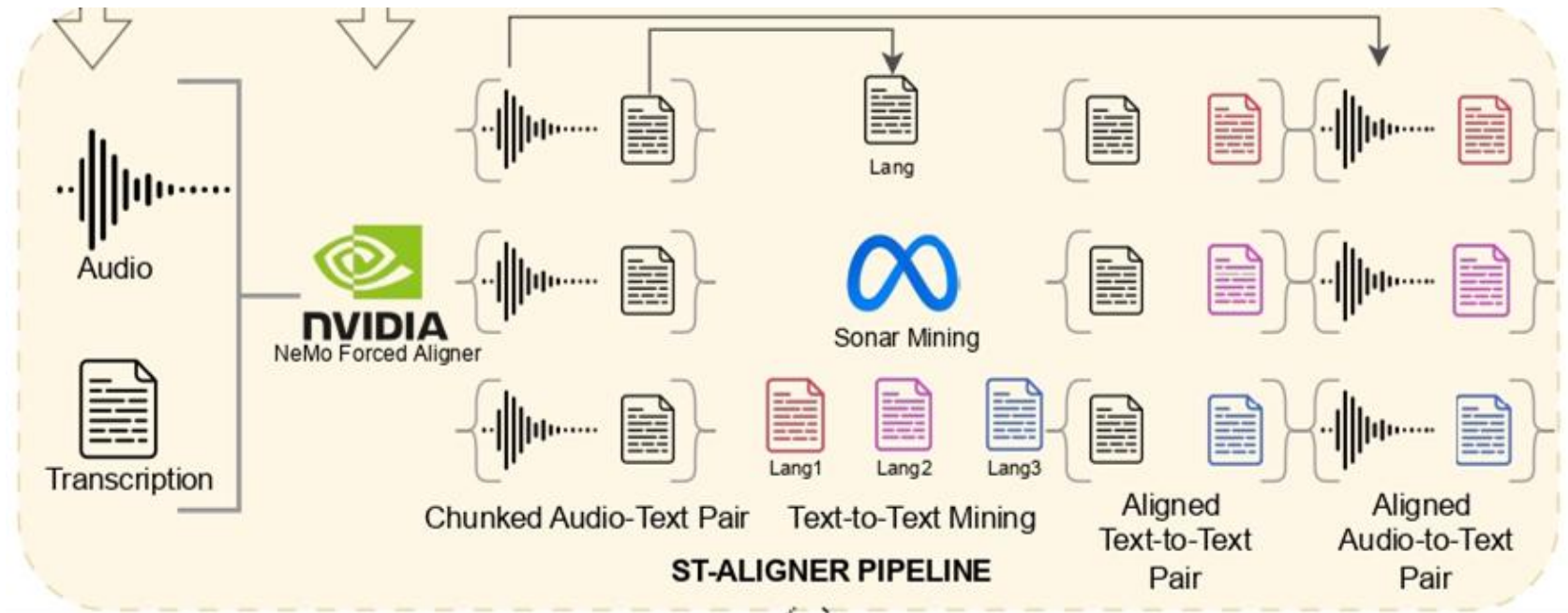
Reduce speech-text mining to Text-text parallel data mining



Align speech-text segments to identify speech/text pairs



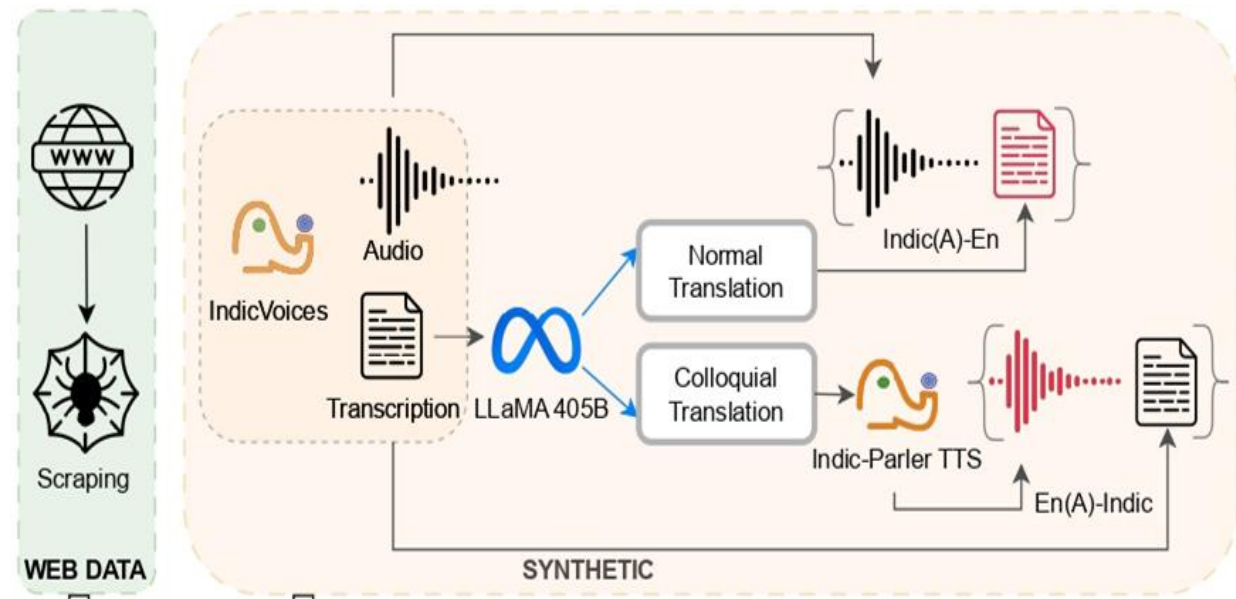
Mine text translation pairs



Synthesizing Speech Translation Data using MT and TTS

Why?

- Available ST data is from limited domains
- Can generate more diverse data using speech and text datasets



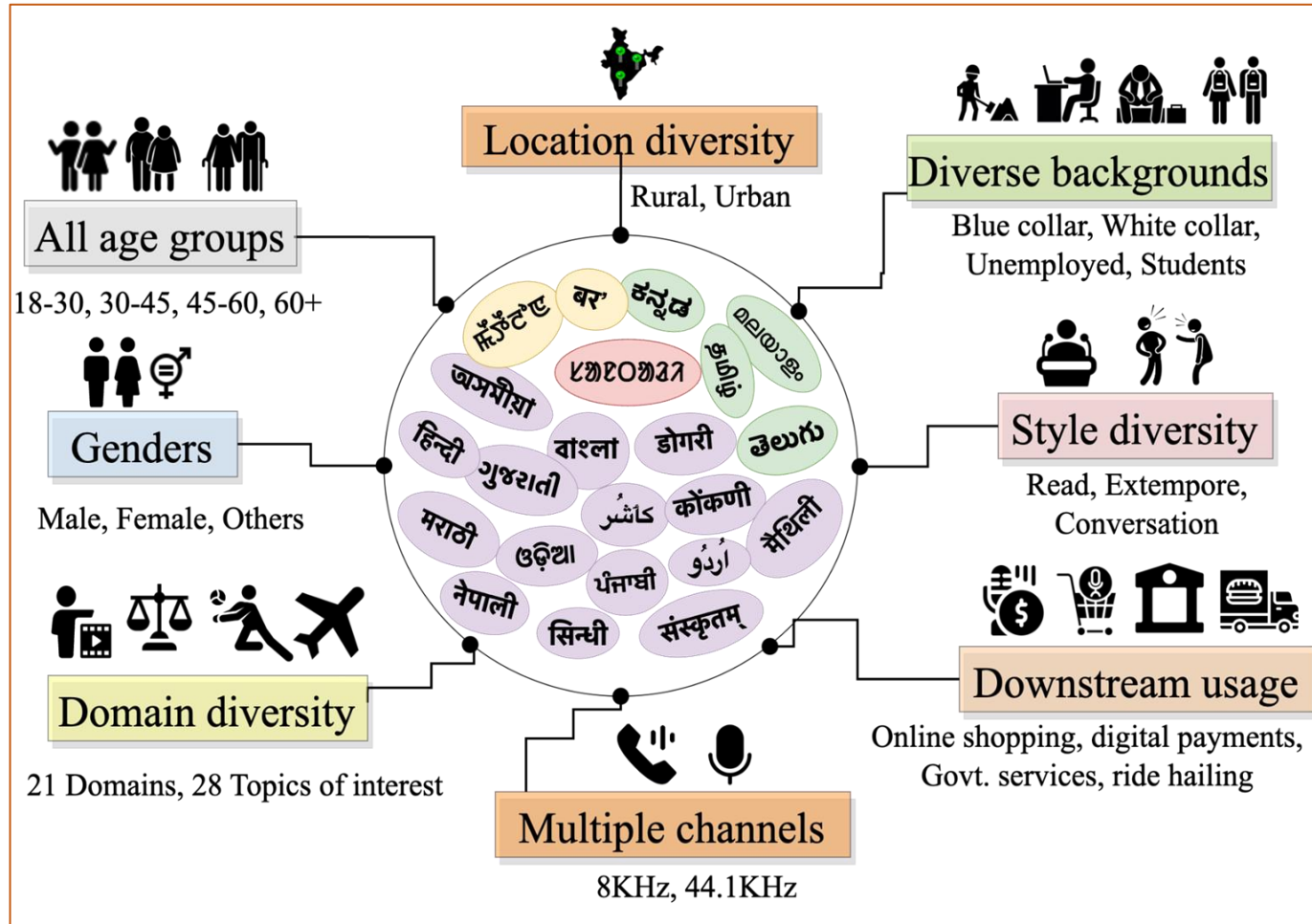
This approach can be generalized to other speech tasks

Collecting ASR data at scale

IndicVoices Project

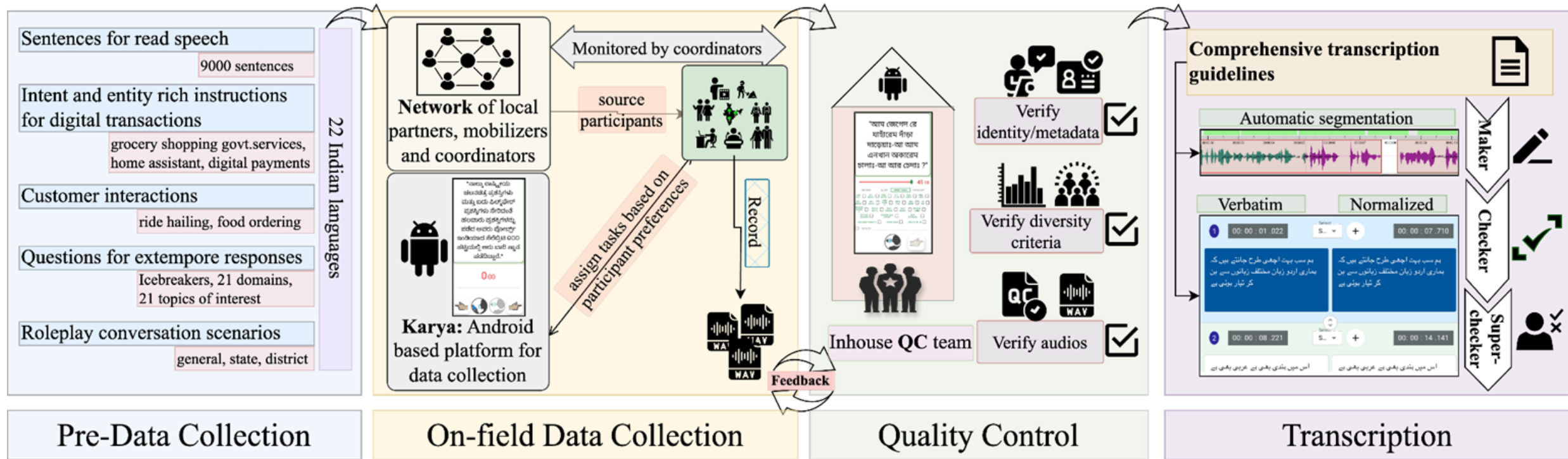
Tahir Javed,, et al.. "Indicvoices: Towards building an inclusive multilingual speech dataset for indian languages." *ACL* (2024).

Defining the wishlist



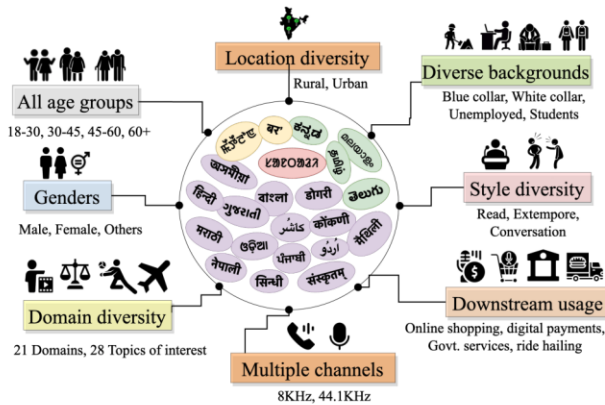
- District wise collection
- Ensuring inclusivity
- Ensuring diversity

Four Key Stages



Three Key Contributions!

Data



IndicVoices

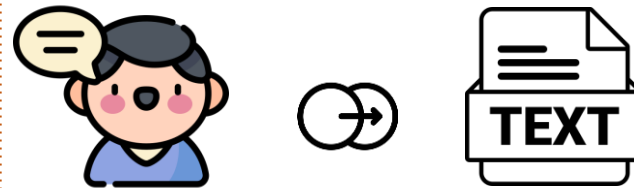
(Goal: **17000** hours)

- **7348** hours (unlabelled)
- **1639** hours (transcribed)
- **22** Languages
- **16237** Speakers

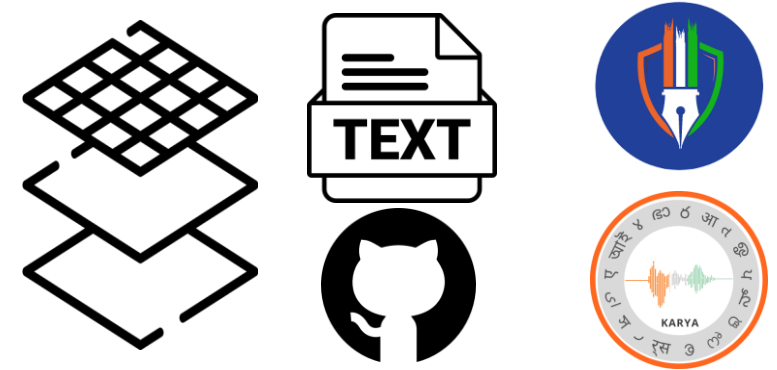
Model

IndicASR

- First to support ASR for all **22** constitutionally recognized languages of India
- Offer lower (Word Error Rate) **WER** than commercial and open source models



Starter kit



Data collection starter kit:

- Collection **blueprint**
- Text resources (**read** commands, **extempore** prompts, **conversational** scenarios)
- Platforms (**Kathbath**, **Shoonya**)

Summary

- Large scale datasets are critical to performance of NLP systems
- Need to harness publicly available datasets and make them available in the public domain
- Innovative ways to mining datasets will help drive progress for many NLP tasks
- Leveraging LLMs to create data for diverse scenarios and tasks
- We need to engage the community for the long tail of languages
- High quality seed data and testsets need to be created with human inputs

Thanks

anoop.kunchukuttan@gmail.com

<http://anoopkunchukuttan.github.io>

<http://huggingface.co/Al4Bharat>

Acknowledgments: All my collaborators, colleagues and students at Al4Bharat and Microsoft

Pre-Data Collection

Sentences for read speech

9000 sentences

Intent and entity rich instructions for digital transactions

grocery shopping govt.services, home assistant, digital payments

Customer interactions

ride hailing, food ordering

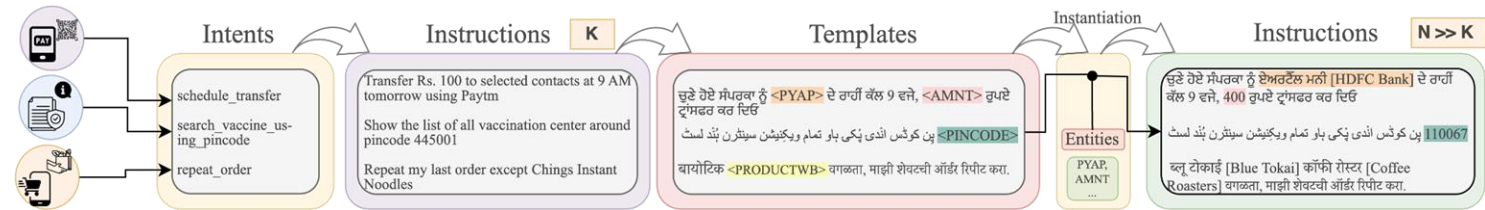
Questions for extempore responses

Icebreakers, 20+ domains, 20+ topics of interest

Roleplay conversation scenarios

general, state, district

22 Indian languages

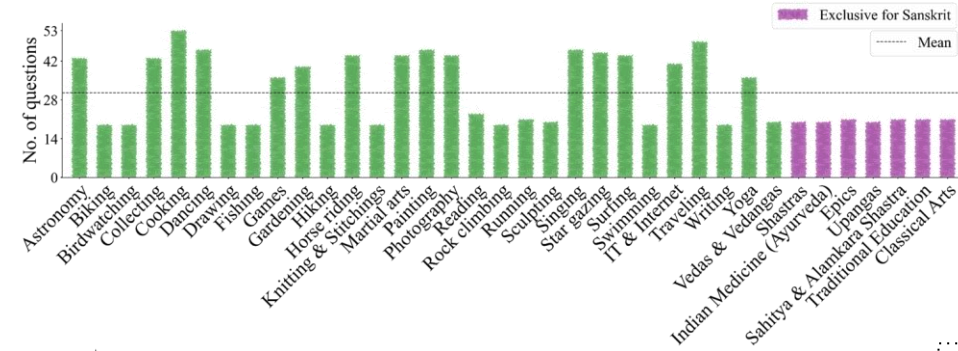


Standardised process to create templates and then instantiate a large number of instructions by replacing entities in the templates

45K+
Sentences

Created

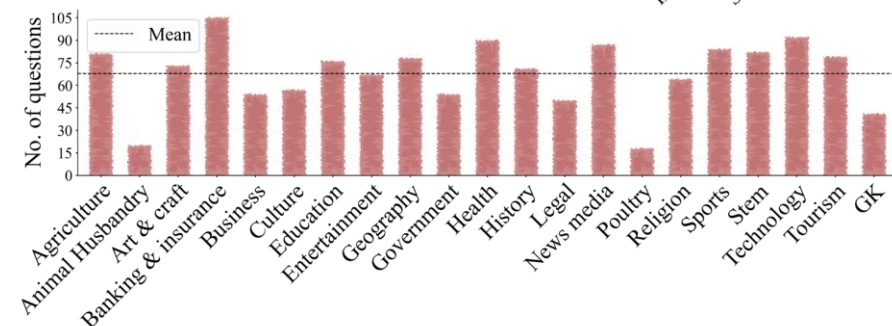
By Experts



20+
Domains

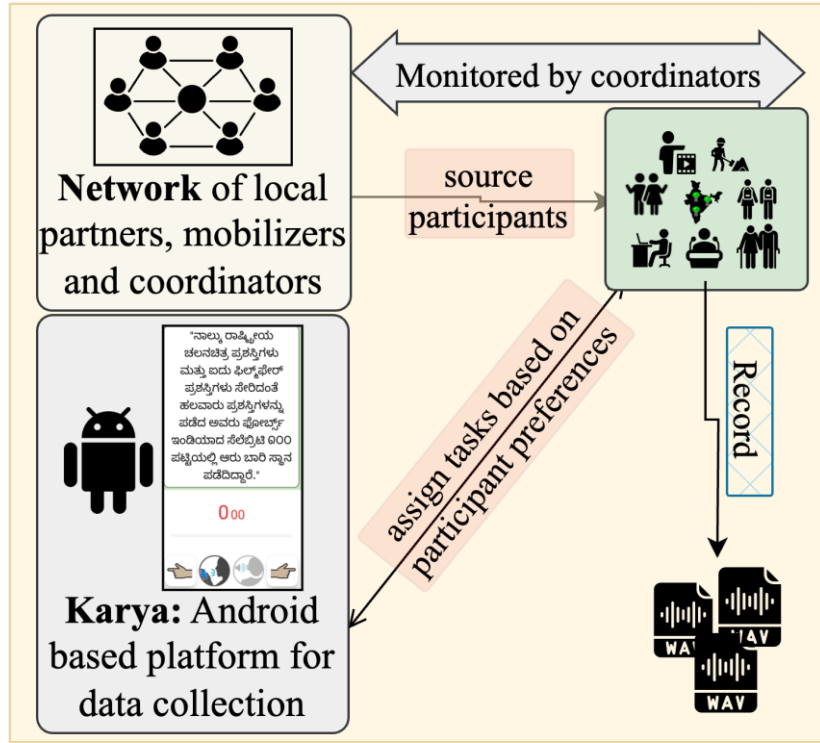
2500+
Extempore Questions

1000+
Conversation scenarios

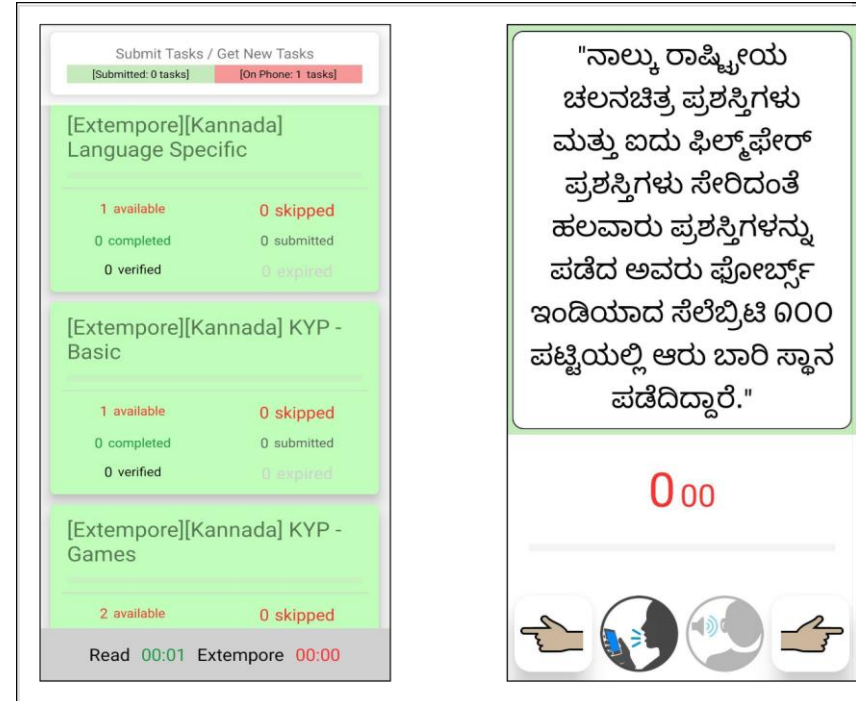


20+
Topics of Interest

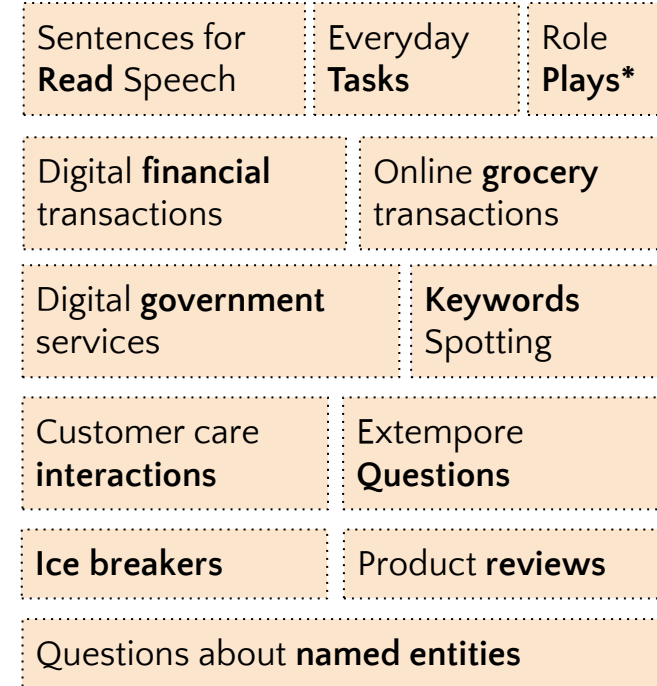
On-field Data Collection



Data collection workflow



Kathbath's Home screen (left), micro-task screen (right)



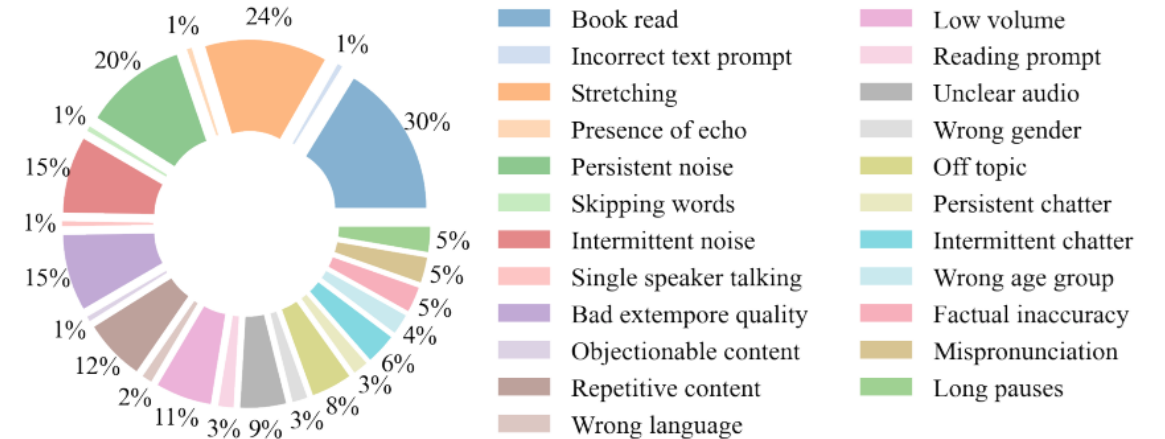
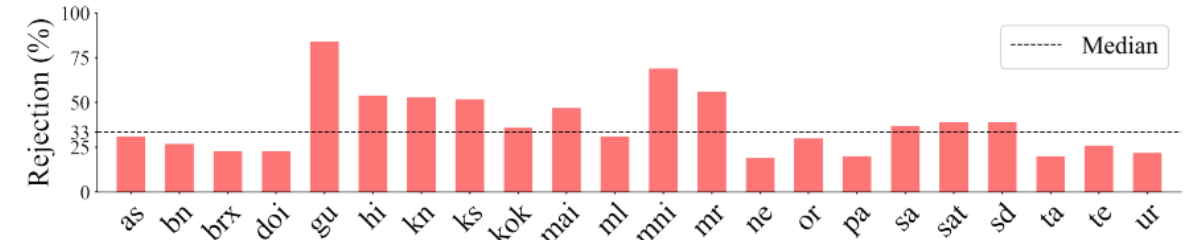
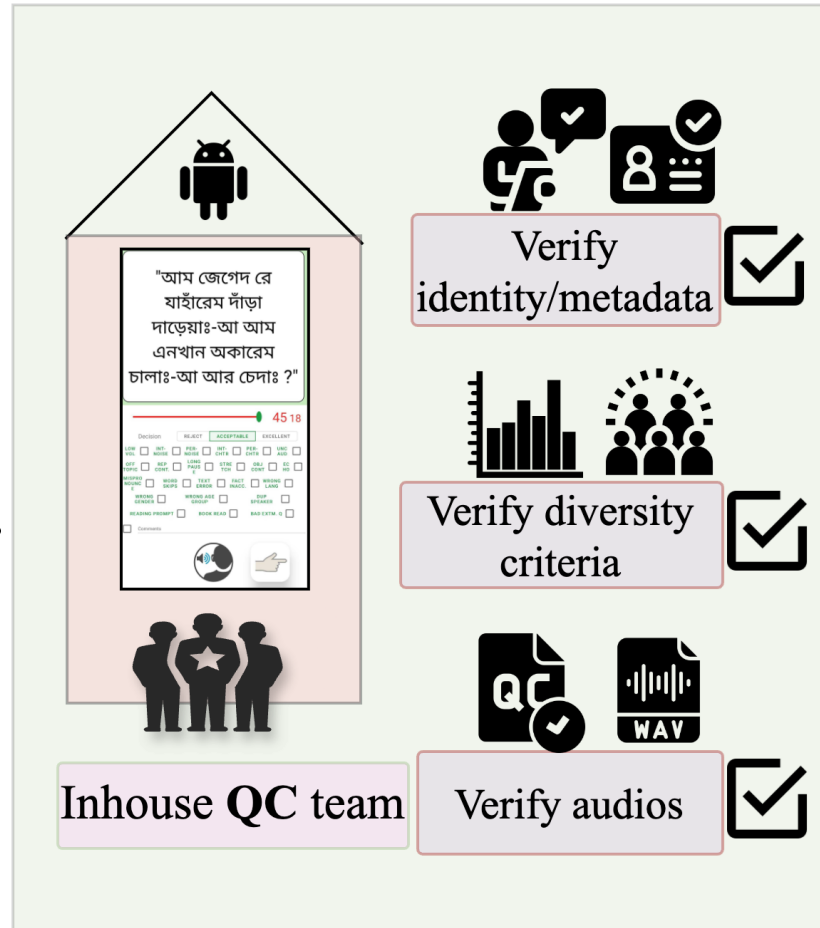
Different tasks supported by Kathbath

Local agencies support seeding of participants, while ensuring **diversity**

Coordinators support with preliminary verification, ensuring **quality**

* Done over Telephony channel

Quality Control



Rejection rate (top) and QC Error categories (bottom)

Transcription

The screenshot displays the Shoonya's Transcription interface. It features two segments of audio transcription. Segment 1 (top) has a duration of 00:00:01.022 to 00:00:07.710 and contains the text: "ہم سب بہت اچھی طرح جانتے ہیں کہ ہماری اردو زبان مختلف زبانوں سے بن کر تیار ہوئی ہے". Segment 2 (bottom) has a duration of 00:00:08.221 to 00:00:14.141 and contains the text: "اس میں ہندی بھی ہے عربی بھی ہے". Below the text, there is a timeline with a waveform and a color-coded bar (green, red, purple) indicating different parts of the audio. The interface includes a 'Select S...' dropdown and a '+' button for each segment.

Shoonya's Transcription interface

Transcription on two levels:

- Verbatim (L1)
- Normalized (L2)

Comprehensive transcription guidelines

Auto Segmentation

