Developing Indian Language Technology for the AI Age

Anoop Kunchukuttan

AI4Bharat

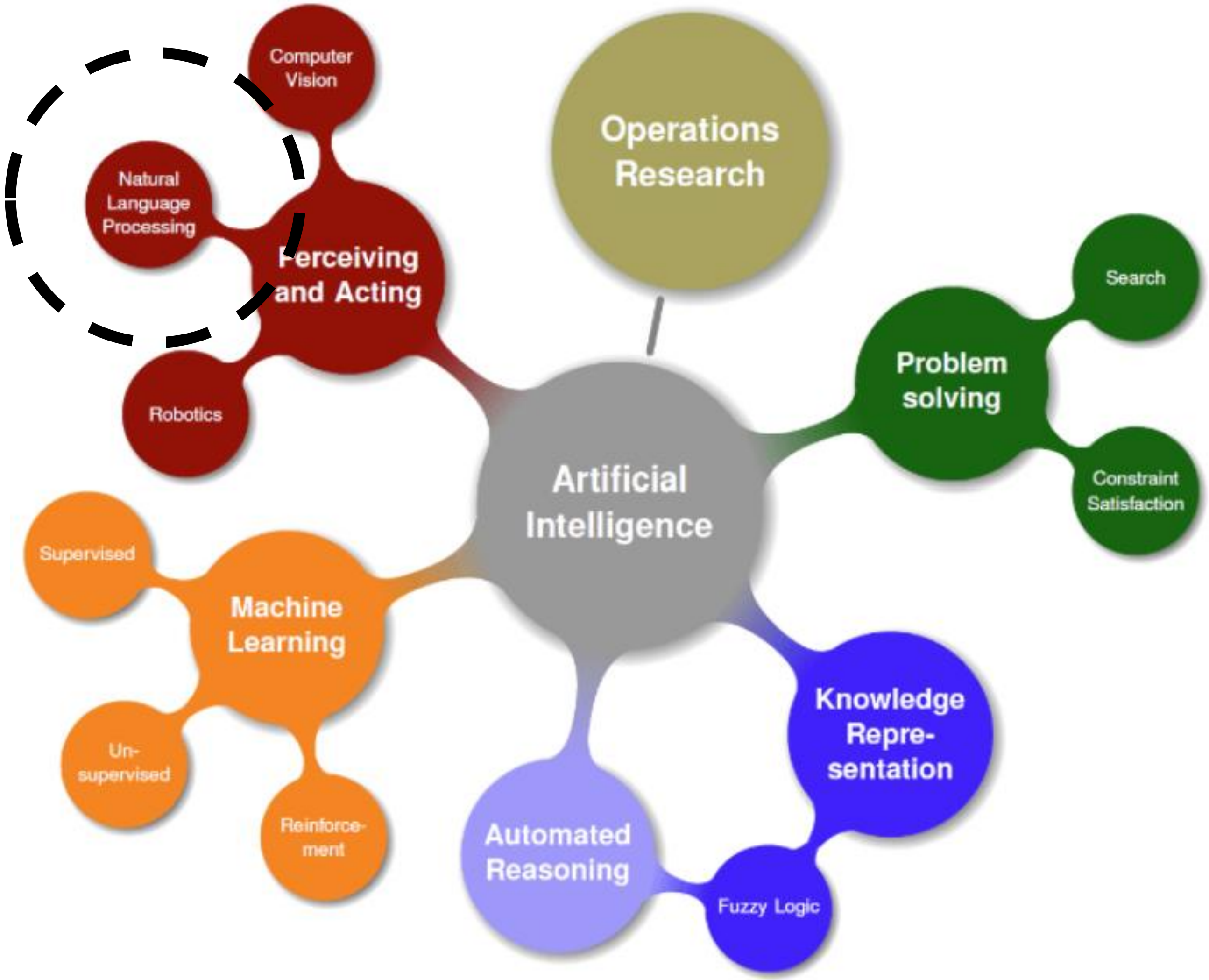
Bharatiya Bhasha Sangam: Exploring Languages in Professional Contexts, EFLU, Hyderabad, 19 Nov 2025

Outline

- **Artificial Intelligence and Language**
- **A Bird's eye-view of Language Technology**
- **Building Indic NLP technology – Pillars**
- **Data Curation at Scale**
- **Multilinguality and Indic language viewpoint**
- **Glimpses of Indic Language Tech Efforts**
- **Summary and Outlook**

What is Artificial Intelligence?

Field of computer science dedicated to creating systems that can perform tasks typically requiring **human intelligence**.



Weak AI Artificial Narrow Intelligence	Strong AI Artificial Narrow Intelligence
Designed and trained for a **single, specific task**. It cannot perform outside its programmed scope.	A theoretical AI that possesses the capacity to understand, learn, and apply knowledge to **solve any problem** a human being can.
All Pervasive – spam filters, recommendation systems, translation systems, facial recognition, and many more …	Science fiction concepts like HAL, C3PO

Can Large Language Models bridge this gap?

Weak AI

Artificial Narrow Intelligence

Designed and trained for a **single, specific task**. It cannot perform outside its programmed scope.

All Pervasive – spam filters, recommendation systems, translation systems, facial recognition, and many more …

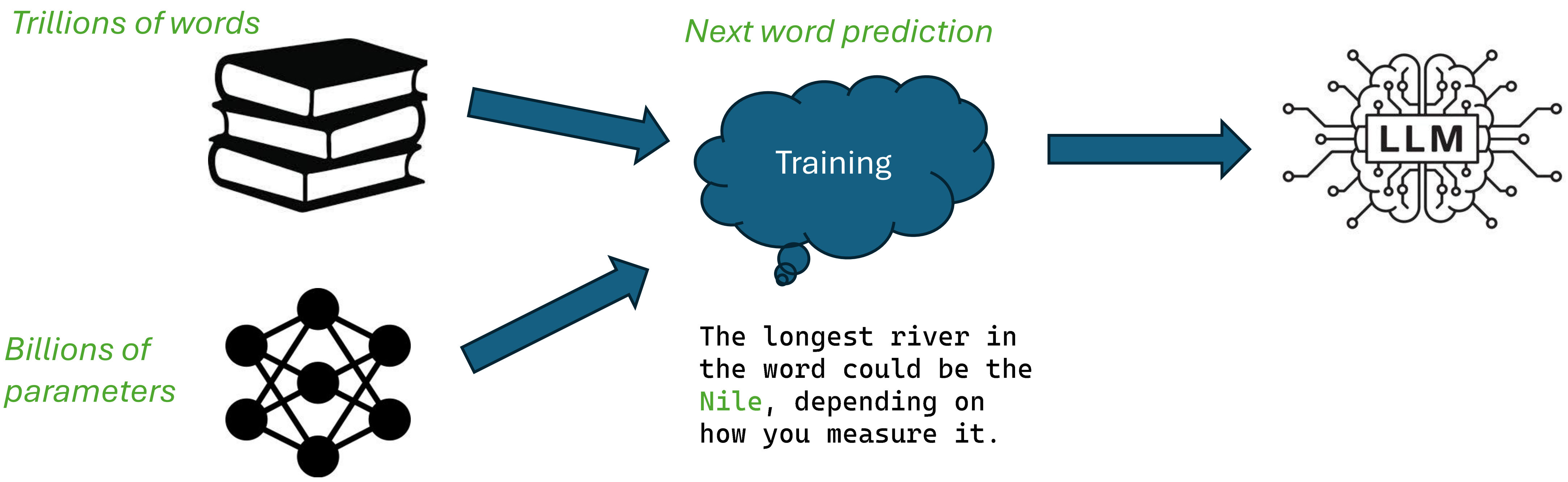
Strong AI

Artificial Narrow Intelligence

A theoretical AI that possesses the capacity to understand, learn, and apply knowledge to **solve any problem** a human being can.

Science fiction concepts like HAL, C3PO

What are Large Language Models?



Basic capability: Auto-complete

The largest river island in the word is ____

Majauli
Australia
Greenland

Learning at scale leads to In-context learning capabilities

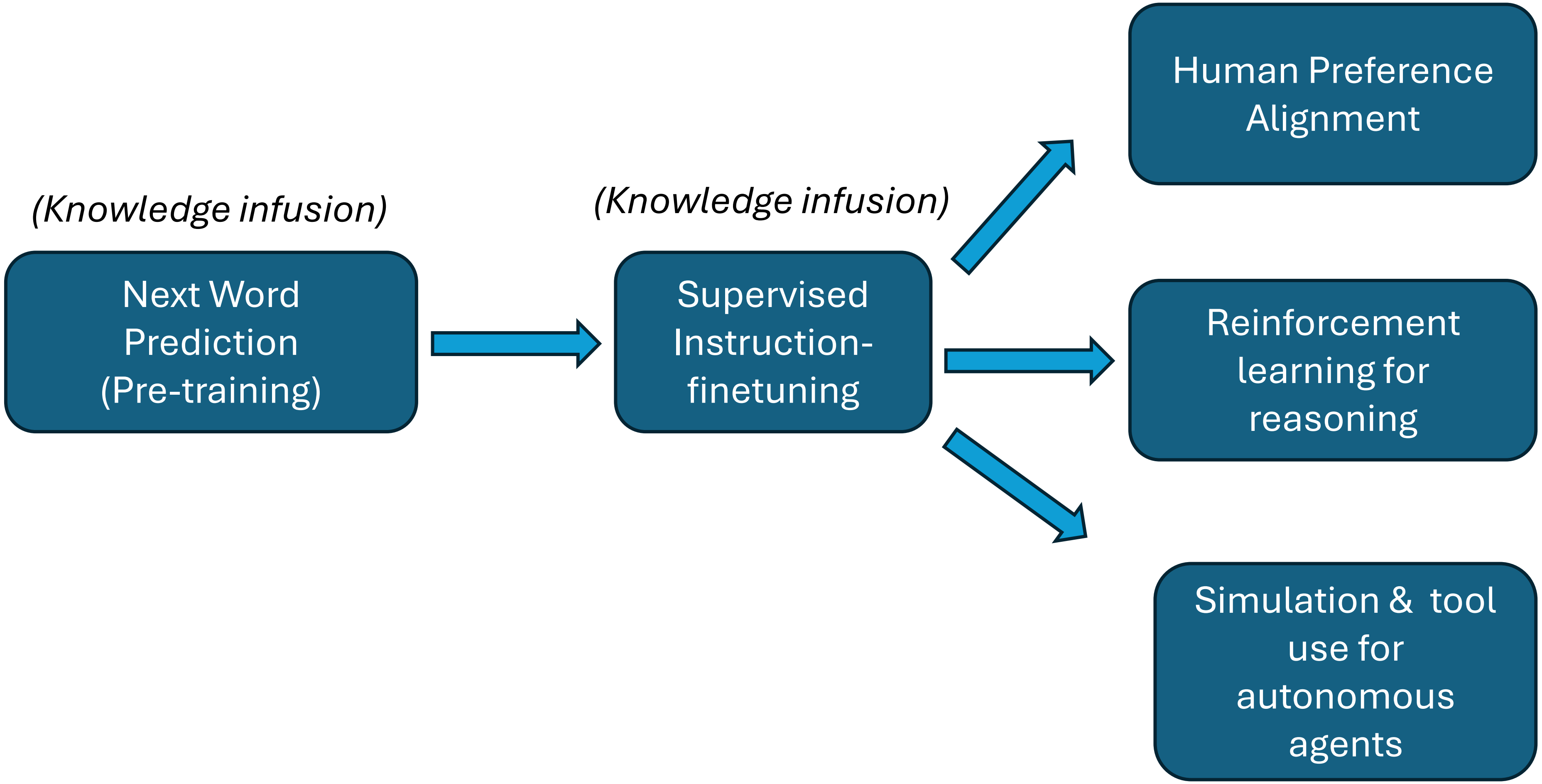
Instruction Tell me the sentiment of this review

Example The movie begins ….. The plot is engaging, thoroughly enjoyable.
The movie is great

Oh, how can such a fine cast produce such a terrible performance….. A total waste of time.
The movie is pathetic

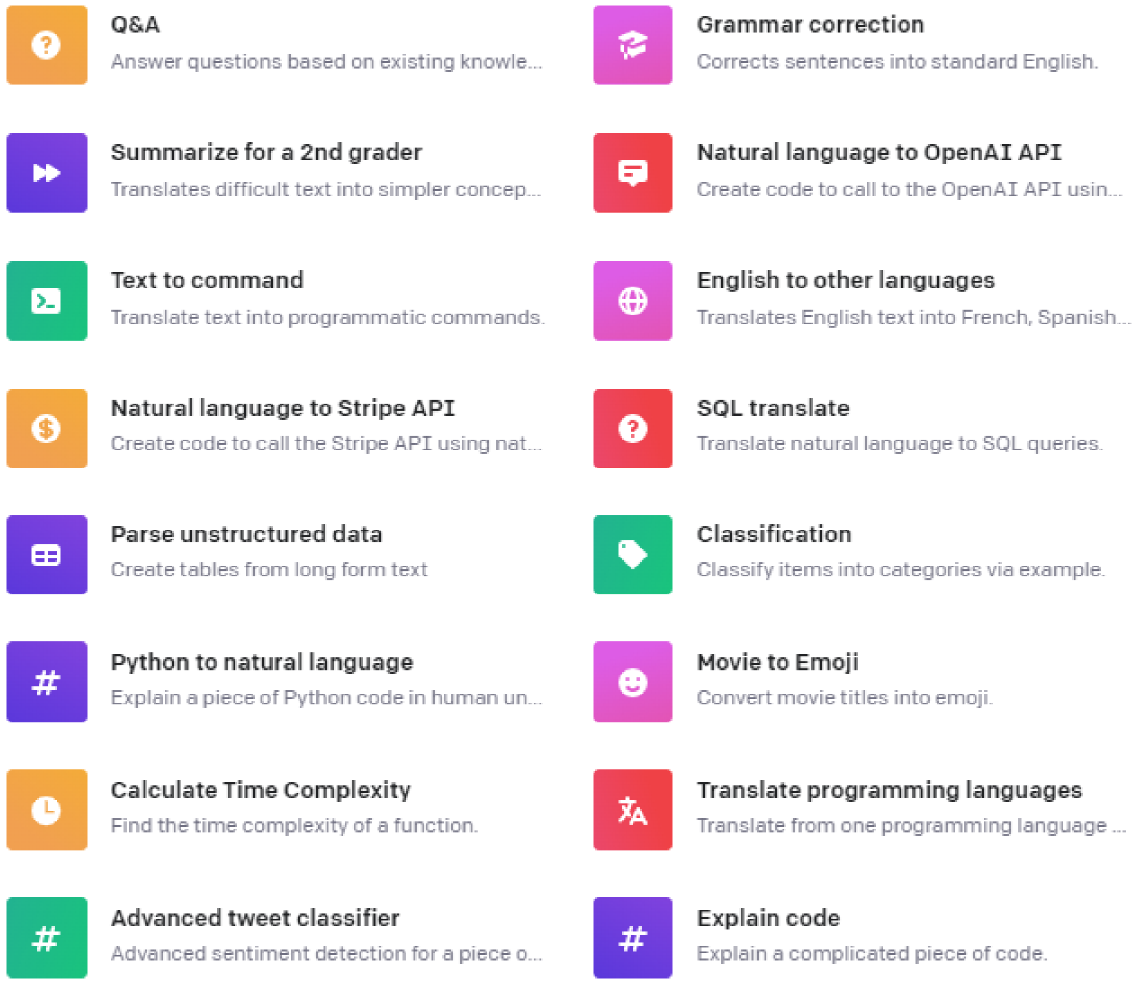
Prompt It is just a rehash of old movies
The movie is <MASK>

How is an LLM trained?



What makes LLMs exciting?

- Inherent knowledge
- Multi-task abilities and task-composition
- Little task-specific training required
- Help in writing , creativity, brainstorming, data generation, explanation, *etc*.
- Reasoning Capabilities



[SparksOfAGI]

https://www.youtube.com/watch?v=qbIk7-JPB2c

Maths

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

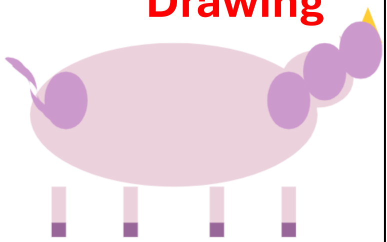
GPT-4:

```
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
```

Drawing

Prompt: Draw a unicorn in TiKZ.

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



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.

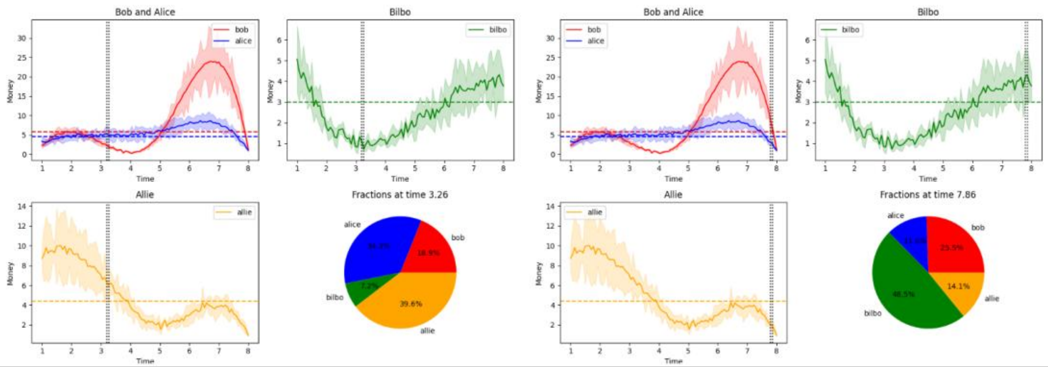
Here is a possible code to generate a pyplot for the following data, with some comments to explain the steps:

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]



GPT-4: **Explanations**

```
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 sec
ond 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:
```

Understanding Programs

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.
```

Reasoning

Why is 31 a prime number?

Reasoned about number 31 being prime for 4 seconds

Alright, let's go through this. A prime number is an integer greater than 1, divisible only by 1 and itself. So why is 31 prime? Its only divisors are 1 and 31. I check for divisibility — it's not divisible by 2 (31/2 = 15.5, remainder 1), not by 3 (31/3 = 10.33, remainder 1), and not divisible by 5 (31/5 = 6.2, remainder 1). Since no other primes divide it, 31 is prime!

Since 31 has no divisors other than 1 and itself, it is prime. To determine this, we check possible divisors up to sqrt(31), approximately 5.6. The divisors to check are 2, 3, and 5, and none divide evenly into 31 — 31/2 = 15.5, 31/3 ≈ 10.33, 31/5 = 6.2. Since 31 isn't divisible by these, it confirms that 31 is prime. The fundamental theorem of arithmetic also supports that 31 cannot be factored further.

[SparksOfAGI]

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)

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

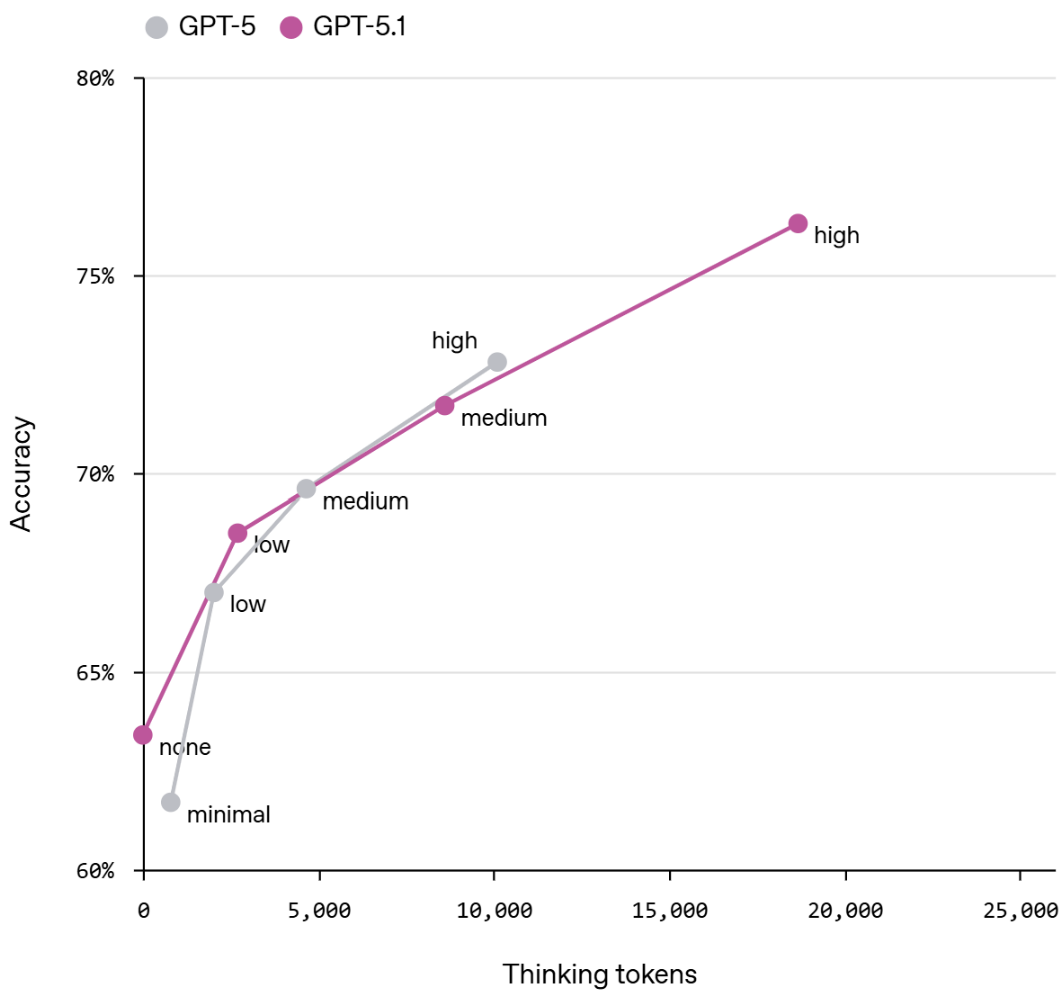
	Qwen3-235B-A22B MoE	Qwen3-32B Dense	OpenAI-o1 2024-12-17	Deepseek-R1	Grok 3 Beta Think	Gemini2.5-Pro	OpenAI-o3-mini Medium
ArenaHard	95.6	93.8	92.1	93.2	-	96.4	89.0
AIME'24	85.7	81.4	74.3	79.8	83.9	92.0	79.6
AIME'25	81.5	72.9	79.2	70.0	77.3	86.7	74.8
LiveCodeBench v5, 2024.10-2025.02	70.7	65.7	63.9	64.3	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
MultiIF 8 Languages	71.9	73.0	48.8	67.7	-	77.8	48.4

Compilation of tasks requiring reasoning skills

(Qwen 3)

Rapid Advances in Coding using AI

- A personal coding assistant
- A debugging partner
- A tutor for new technologies
- A code reviewer
- A productivity multiplier
- A rapid prototyping engine



Improved Code Capabilities → Rapid Advances in Problem Solving Across Domains

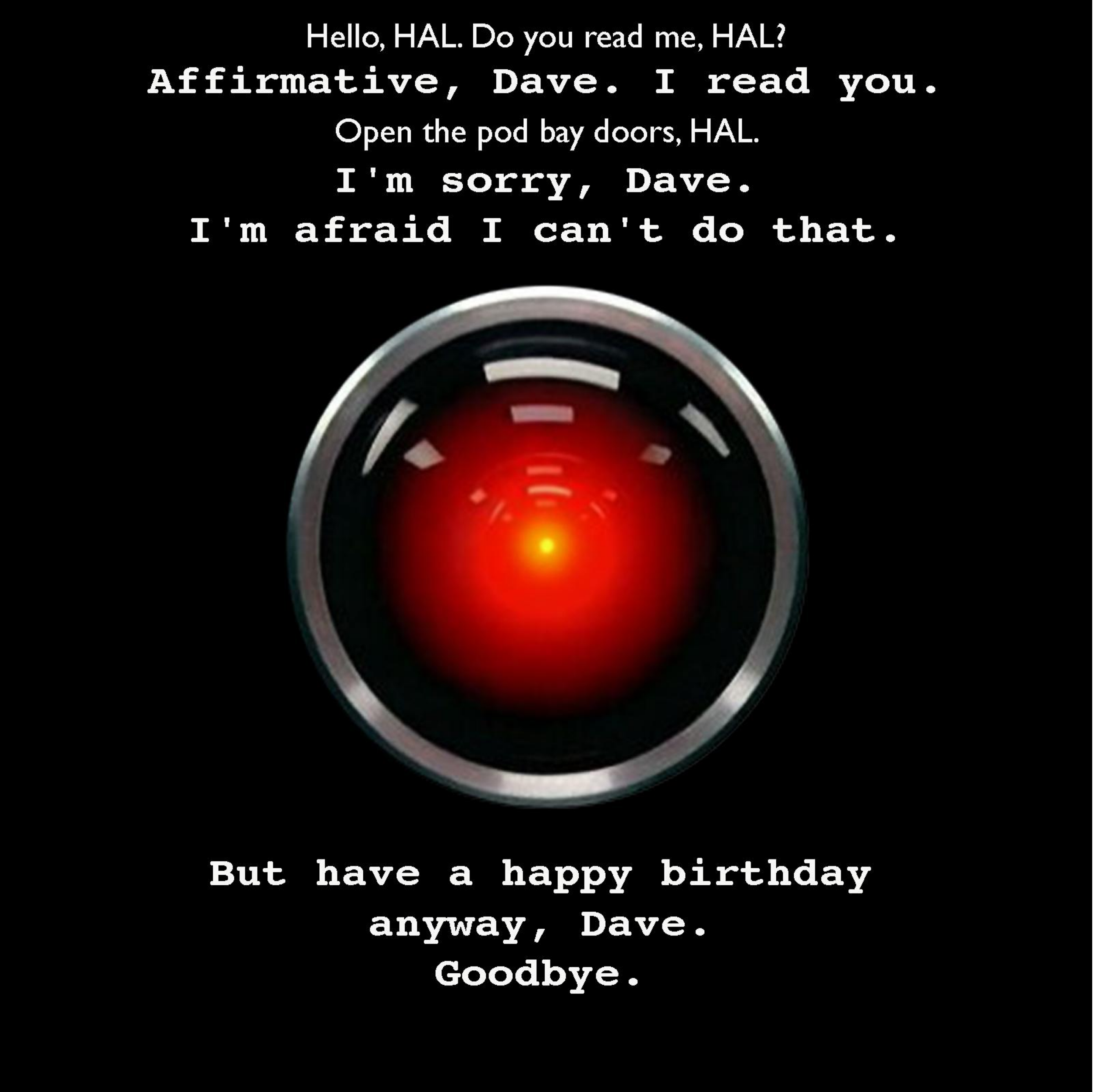
Language is not just another area of Artificial Intelligence

Natural Language is the “language” of today’s most sophisticated LLMs

Knowledge Representation

Reasoning

Planning



Natural Language Processing deals with the interaction between computers and humans using natural language.

An intelligent agent like HAL can do:

- *Natural Language Understanding*
- *Natural Language Generation*

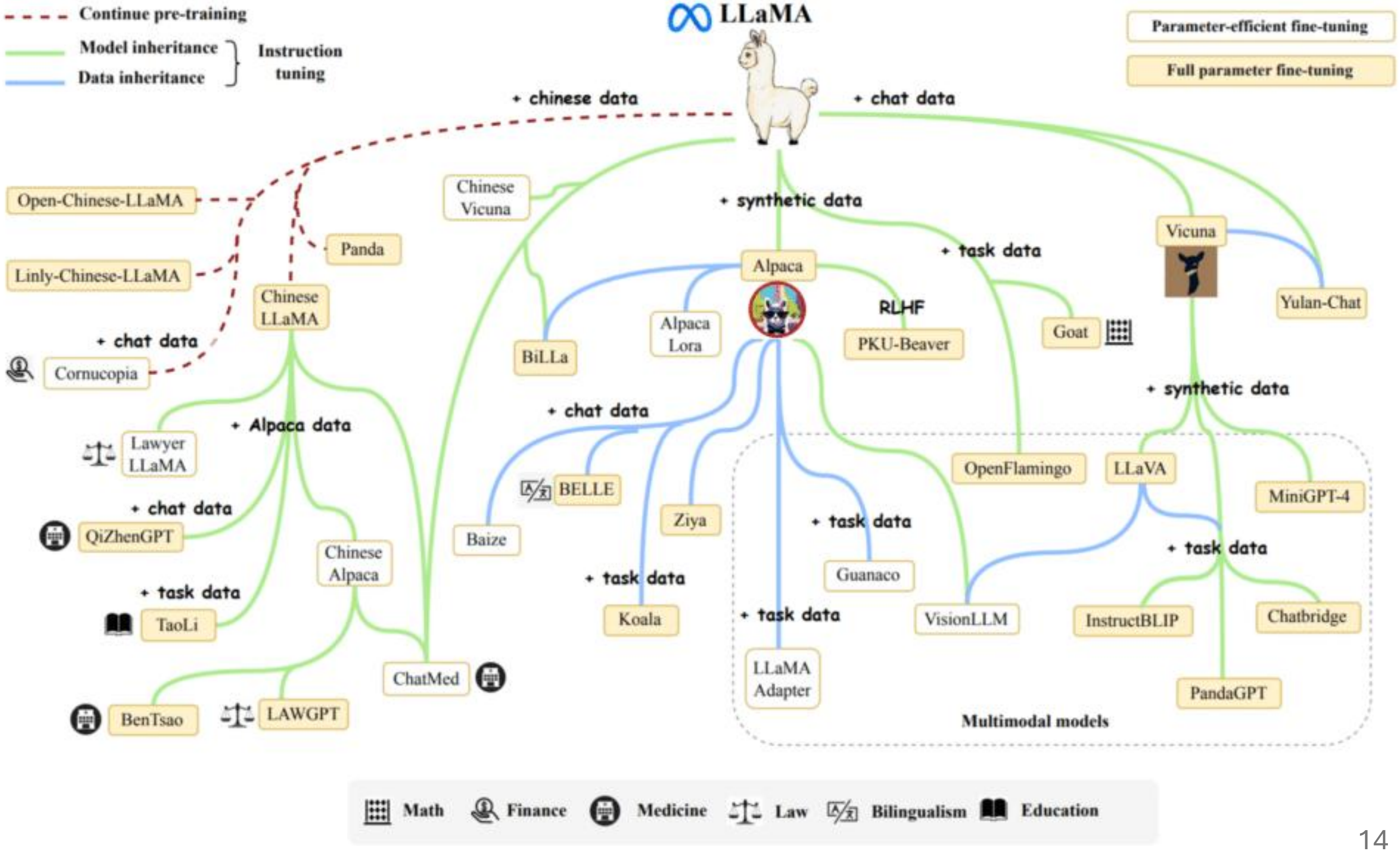
Many other useful applications

- *Text Classification*
- *Spelling Correction*
- *Grammar Checking*
- *Essay Scoring*
- *Machine Translation*

Explosion of LLMs ... but mostly limited to English

Gemini
ANTHROP\C
OpenAI
MISTRAL AI_
cohere
Grok

Gemma
Phi
Qwen
deepseek



Benefits of LLMs are mostly limited to English

Results on XNLI

Language	Cat.	ChatGPT (en)	ChatGPT (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

Language	Cat.	ChatGPT(en) EM	ChatGPT(en) 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 XQuad QnA

Model	EN	AVG
Qwen2.5-32B-Instruct	38.43	29.41
Gemma3-27B-IT	50.55	44.88
QwQ-32B	79.43	74.69
Deepseek-R1	78.81	75.72
o3-mini	82.18	79.90

Performance on MMATH

Lang.	ChatGPT BLEU	ChatGPT chrF	NLLB BLEU	NLLB 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

Results on X-CSQA

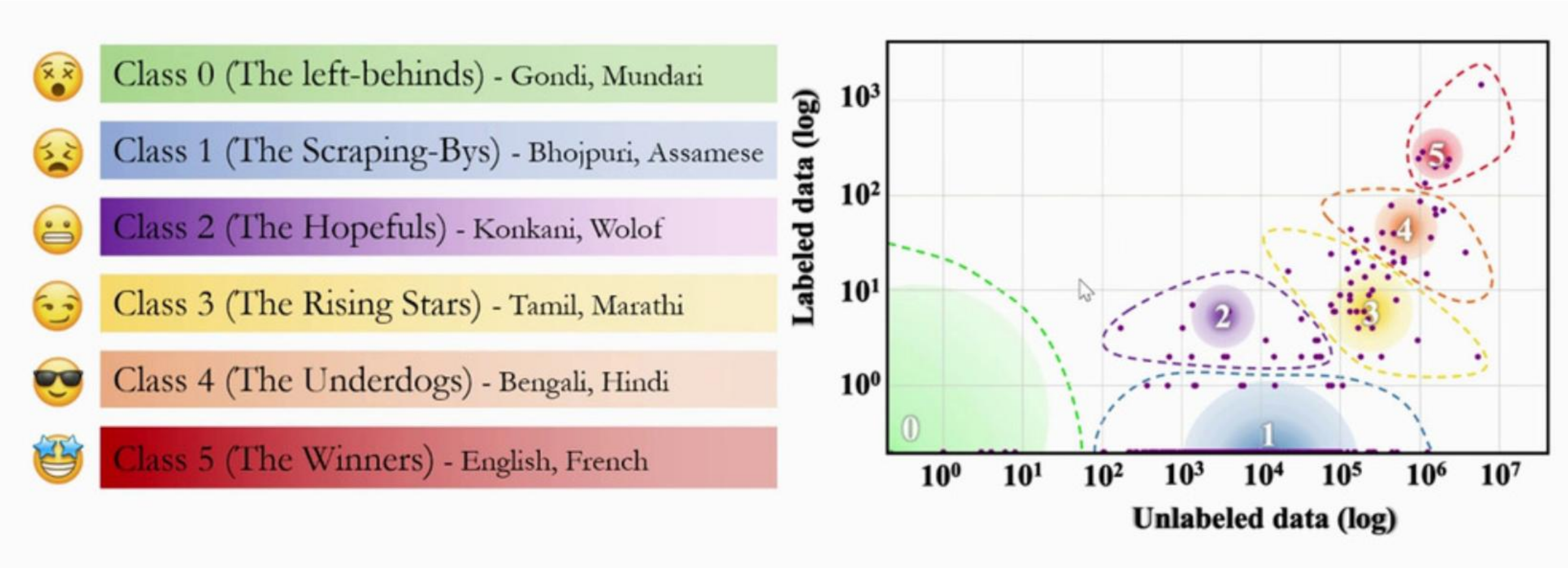
Language	Code	Cat.	ChatGPT (en)	ChatGPT (tgt)
English	en	H	75.0	75.0
Russian	ru	H	50.2	53.5
German	de	H	52.6	61.0
Chinese	zh	H	50.2	42.5
Japanese	jp	H	41.9	43.0
French	fr	H	50.5	61.7
Spanish	es	H	53.3	62.5
Italy	it	H	50.6	55.9
Dutch	nl	H	52.9	60.4
Polish	pl	H	35.2	51.1
Portugese	pt	H	49.5	59.2
Vietnamese	vi	H	42.3	47.9
Arabic	ar	M	49.4	47.3
Hindi	hi	M	41.1	38.6
Urdu	ur	L	34.7	24.5
Swahili	sw	X	35.6	46.6
Average			47.8	51.9

- 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,MMATH]

Disparity in linguistic resources has always been an issue for NLP

Wikipedia/CommonCrawl data as a proxy for monolingual data availability



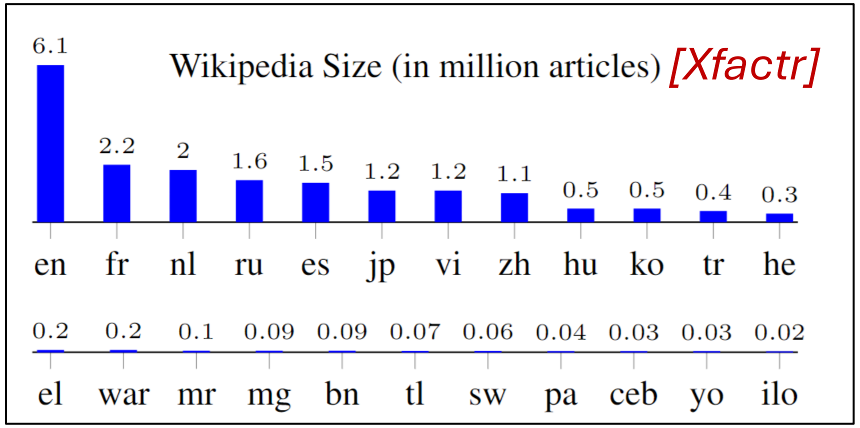
How do we bring the state-of-the-art NLP solutions to all languages?

Can we train such large models for all languages?

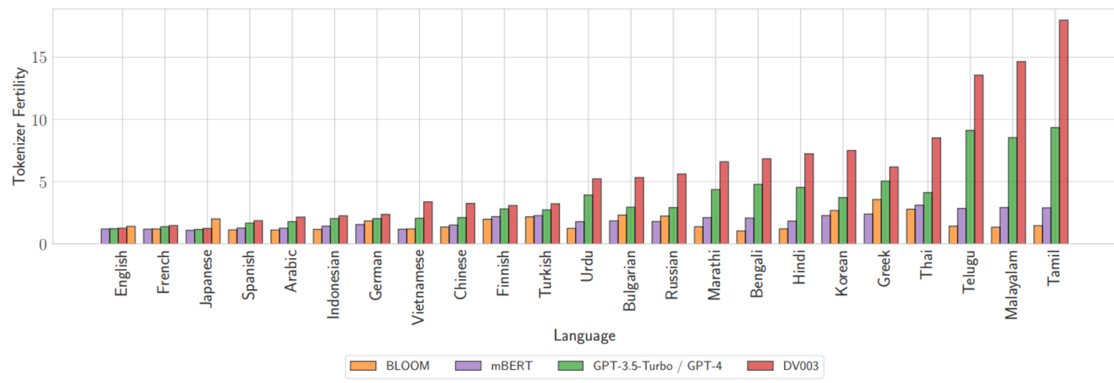
Joshi et al. The State and Fate of Linguistic Diversity and Inclusion in the NLP World. ACL 2023.

Why do LLMs lag behind for other languages?

- Lack of
  - Pre-training data
  - Token representation
  - Instruction tuning data
  - Human preference data
  - Reasoning data
- Inability to transfer from English
- Limitations of Translation Solution



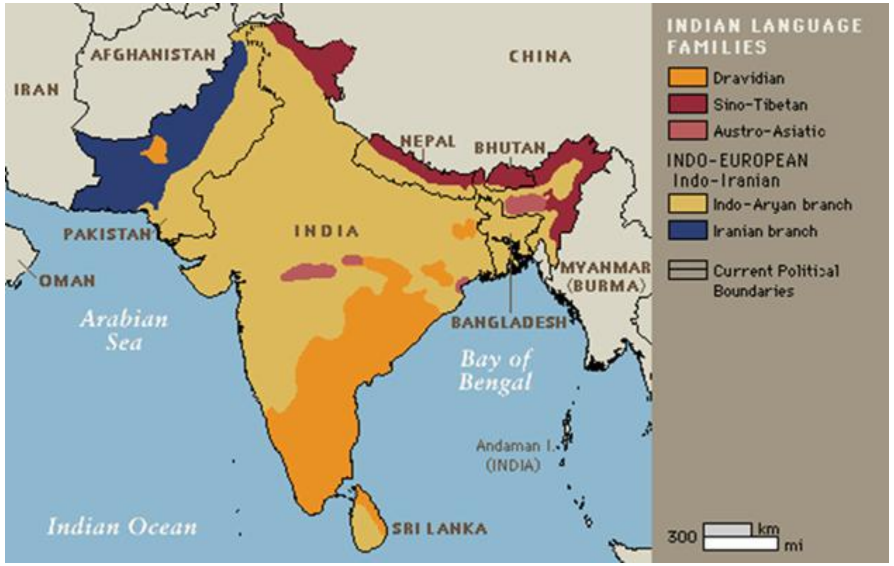
Most LLMs trained on <10% non-English data



Fertility → number of tokens per word
High fertility → low-efficiency, suboptimal representations

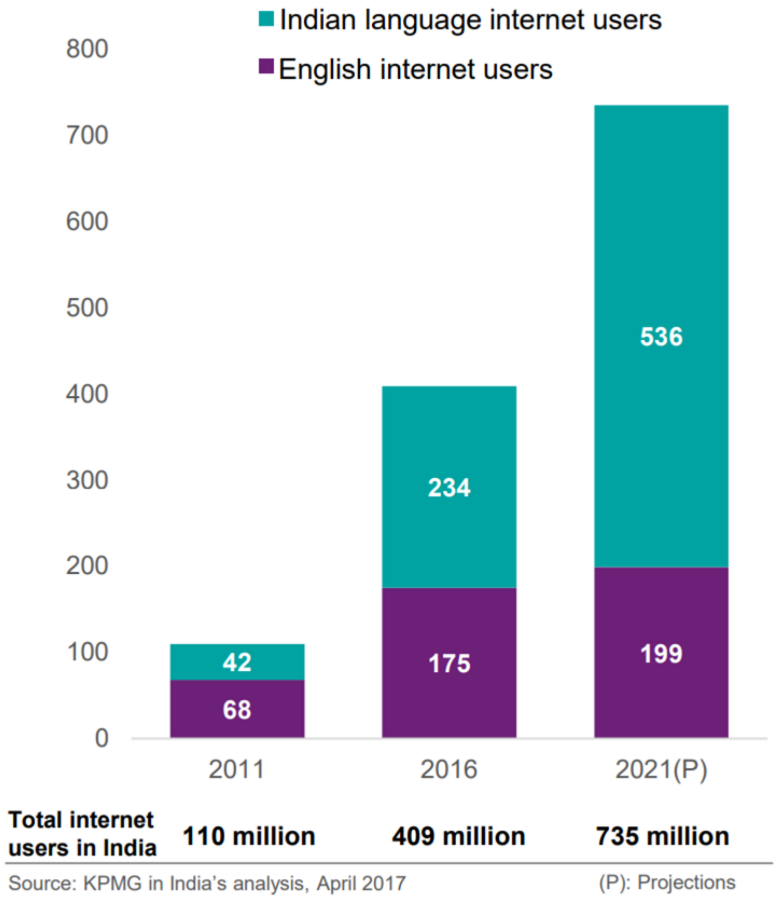
[BUFFET, MEGA, ChatGptMT]

Usage and Diversity of Indian Languages



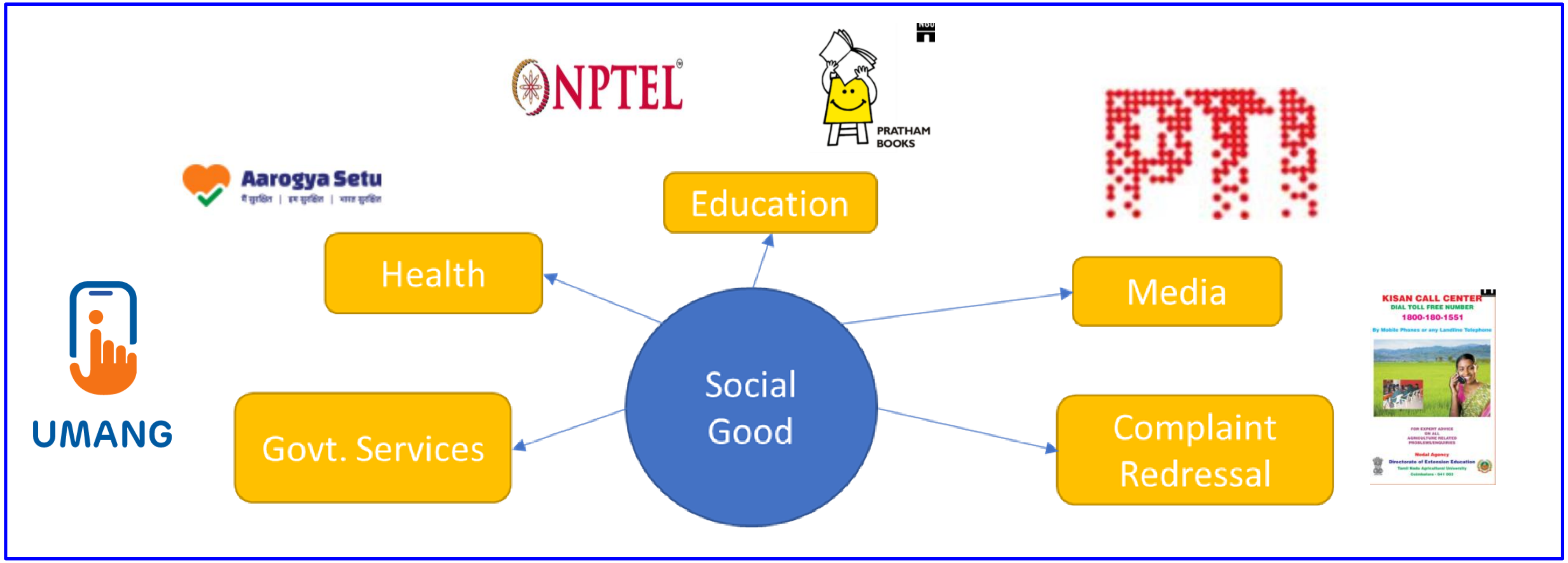
- *4 major language families*
- *22 scheduled languages*
- *125 million English speakers*
- *8 languages in the world's top 20 languages*
- *30 languages with more than 1 million speakers*

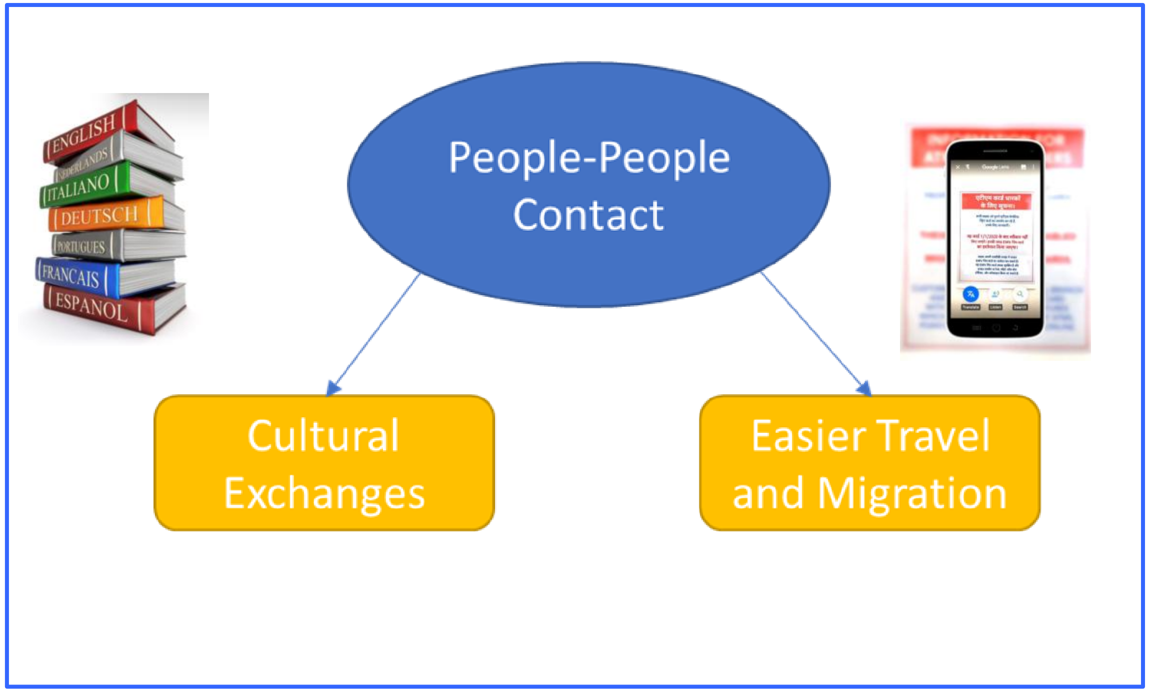
Sources: Wikipedia, Census of India 2011

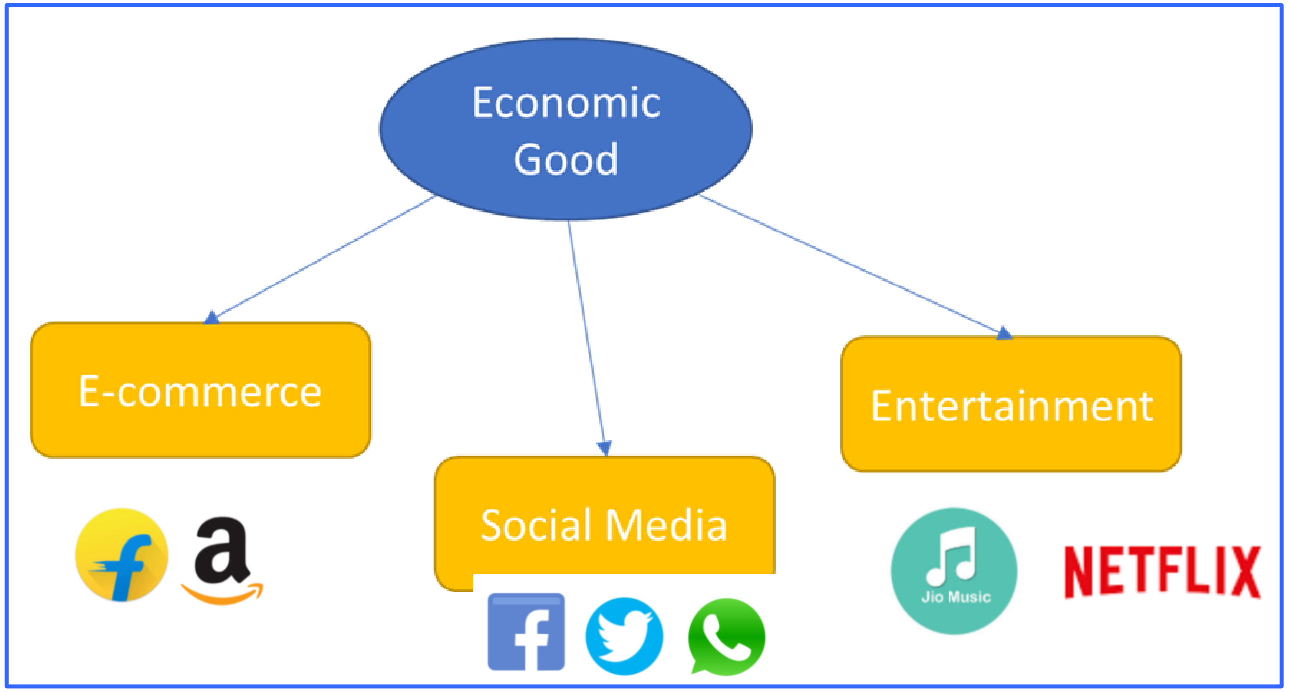


Internet User Base in India (in million)

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


Social Good
Health
Education
Media
Govt. Services
Complaint Redressal
Aarogya Setu
NPTEL
PRATHAM BOOKS
UMANG
KISAN CALL CENTER
DIAL TOLL FREE NUMBER
1800-180-1551

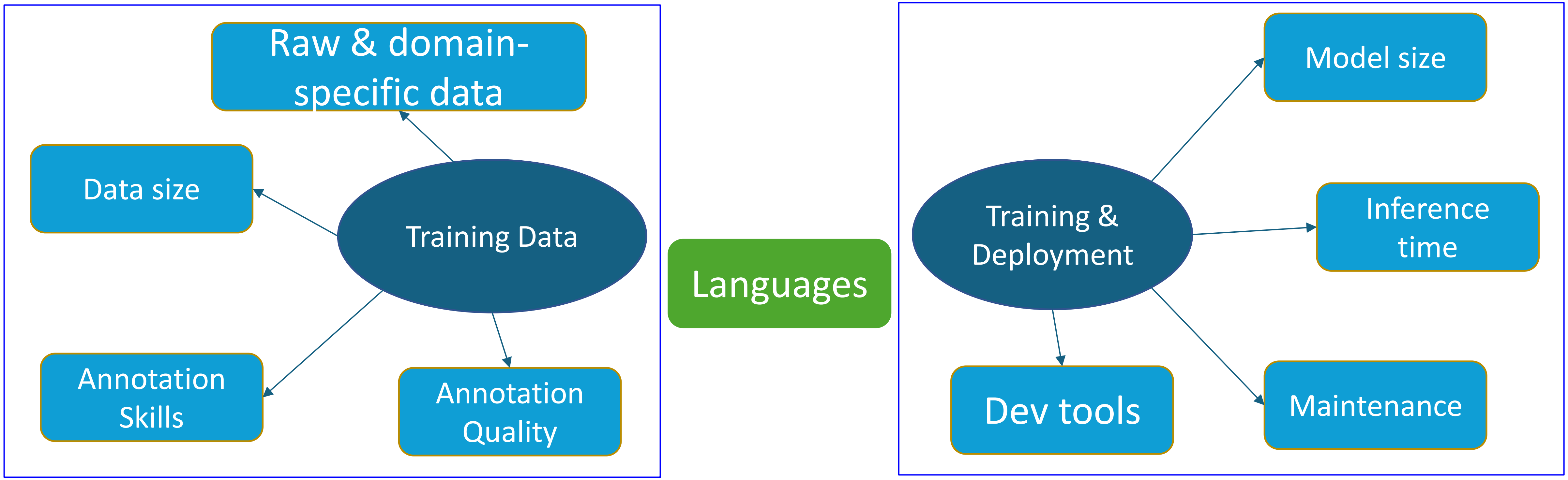

People-People Contact
Cultural Exchanges
Easier Travel and Migration
ENGLISH
ITALIANO
DEUTSCH
FRANCAIS
ESPANOL


Economic Good
E-commerce
Social Media
Entertainment
Jio Music
NETFLIX

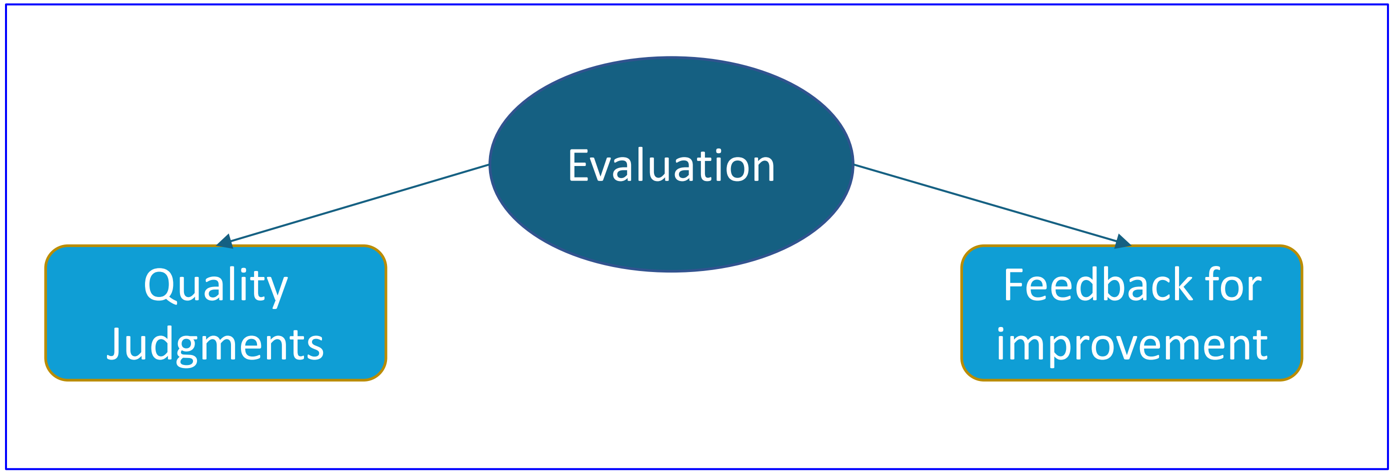
We are faced with a huge data skew

Raw Text Corpora	*Wikipedia articles*	*English*	***6m***
		Hindi	***150k***
Parallel Corpora	*Sentence pairs*	*En-fr (OPUS)*	***500m***
		En-hi (IITB)	***1.5m***
NER Corpora	*Tokens*	*en (CoNLL 2003)*	***200k***
		hi (FIRE)	***40k***
QA	*Question-Answer Pairs*	*en (SQuAD 1.1)*	***100k***
		hi (MMQA)	***4.6k***

Scalability Challenges for NLP solutions



Effort and cost increase as languages increase



How do build language technology solutions for Indian languages

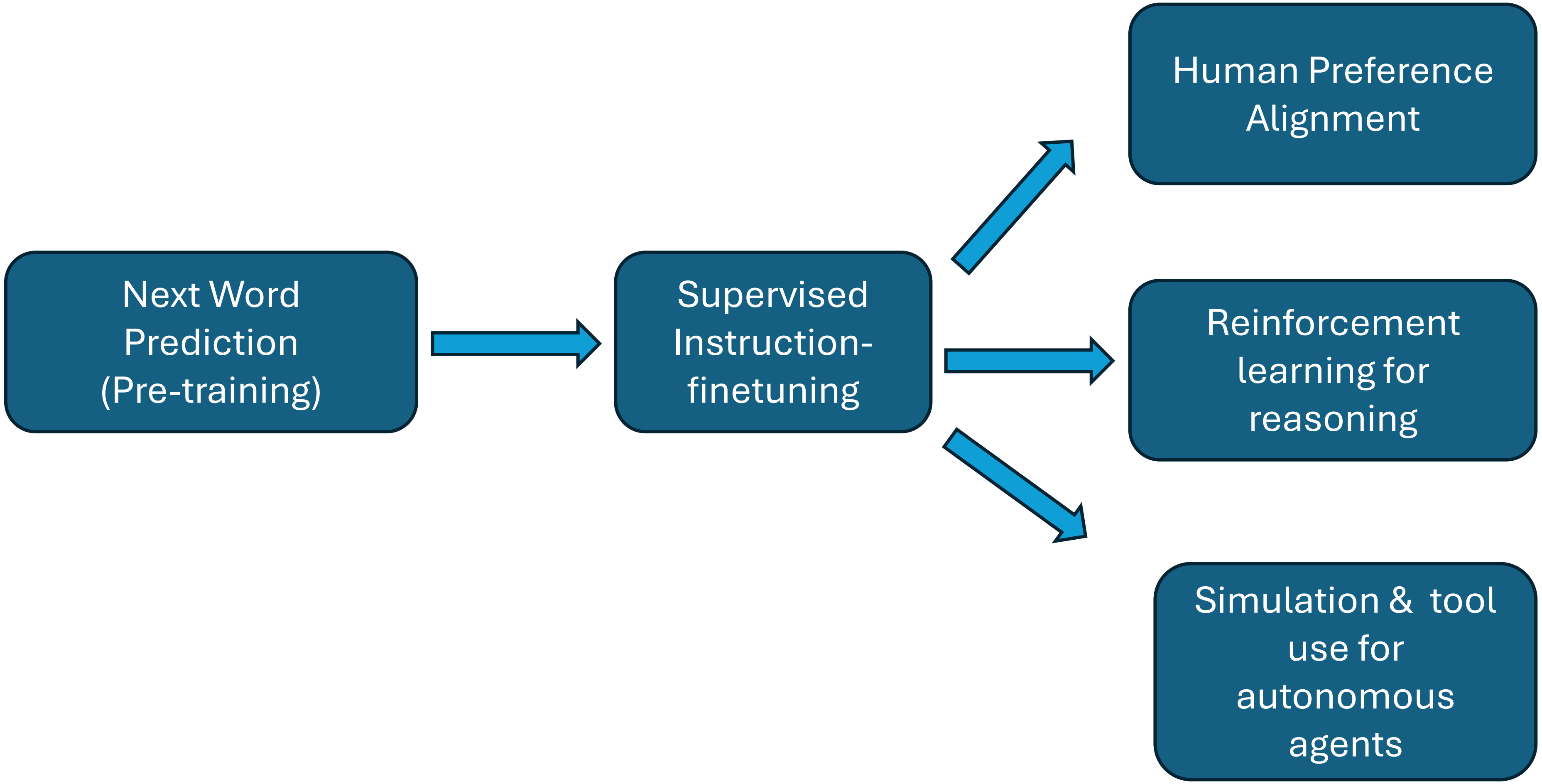
that are of high-quality &

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

In the context of these stages



Collaborative AI empowers everyone to innovate

Open-weights

Open-source

Industry-academia collaborations

- ***Faster Innovation***
- ***Customization and Flexibility***
- ***Scientific Progress and Reproducibility***
- ***Educational Value***
- ***Reduced Costs and Lower Barriers to Entry***
- ***Transparency and trust***

Datasets and models are fundamental infrastructure, they need to be open

AI4Bharat

Sharing compute resources with aligned goals

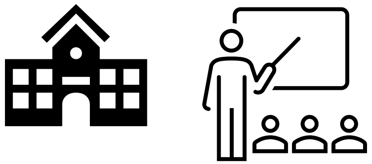
Compute
Cloud

While we develop models and businesses, we must share the knowledge to setup a virtuous innovation cycle

Ai2

OLMo 2
Tülu 3
Molmo

deepseek
R1



Byte-Pair encoding
Direct Preference Optimization

Data Curation at Scale

We need diverse types of data and innovative methods and processes to collect them

- **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

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

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

Publicly Multilingual corpora are good starting points

Major Corpora

mC4, CC100

Wikipedia

OSCAR

CulturaX

MADLAD

Glot-500

MALA-500

✓ Large-scale, Document-level Datasets

? High Quality Documents

? Wide coverage of topics

✗ Representation of culture-specific data, native literature

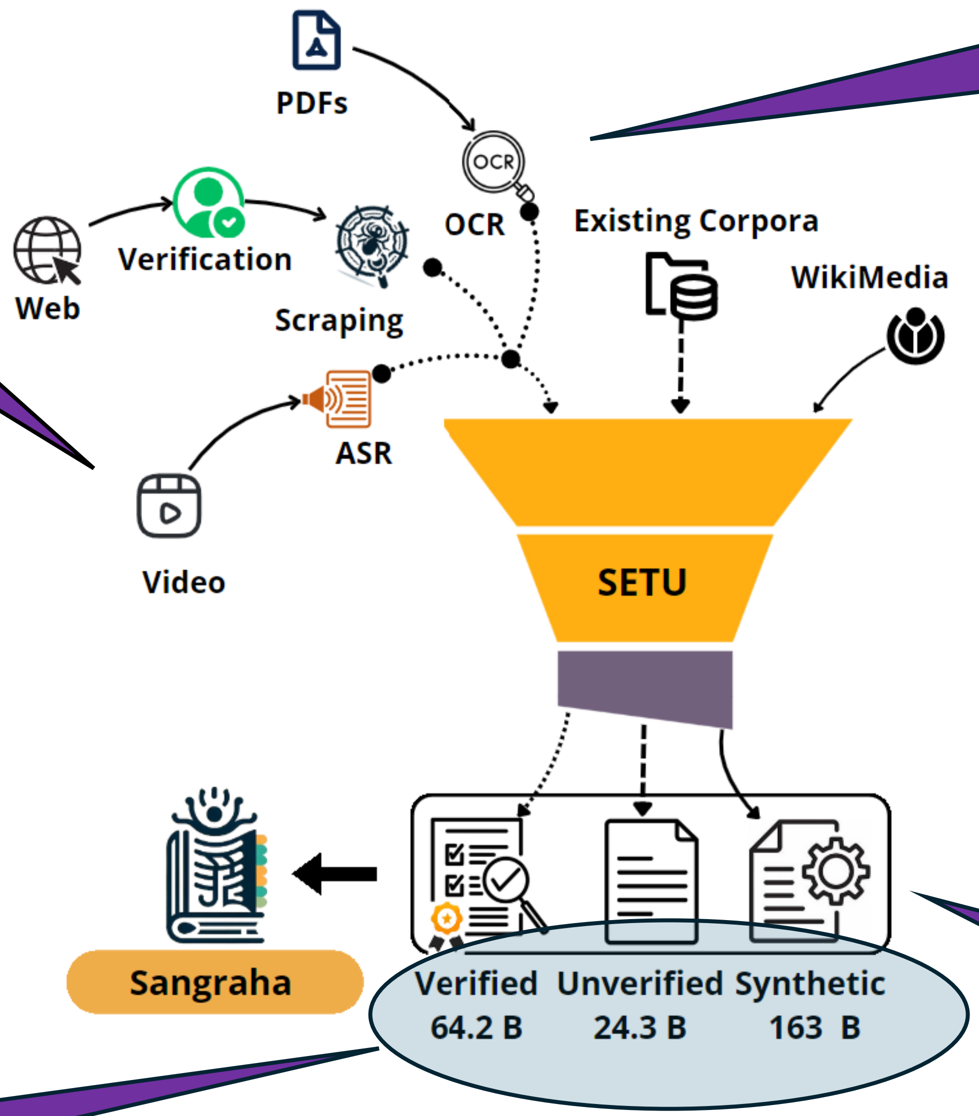
✗ Capture data in different modalities and genres

✗ Data to Help Cross-lingual transfer with English

Build custom language (group) specific collections to address gaps

Data from different modalities – audio/OCR

OCR to unlock diverse regional and cultural content in PDFs



SANGRAHA

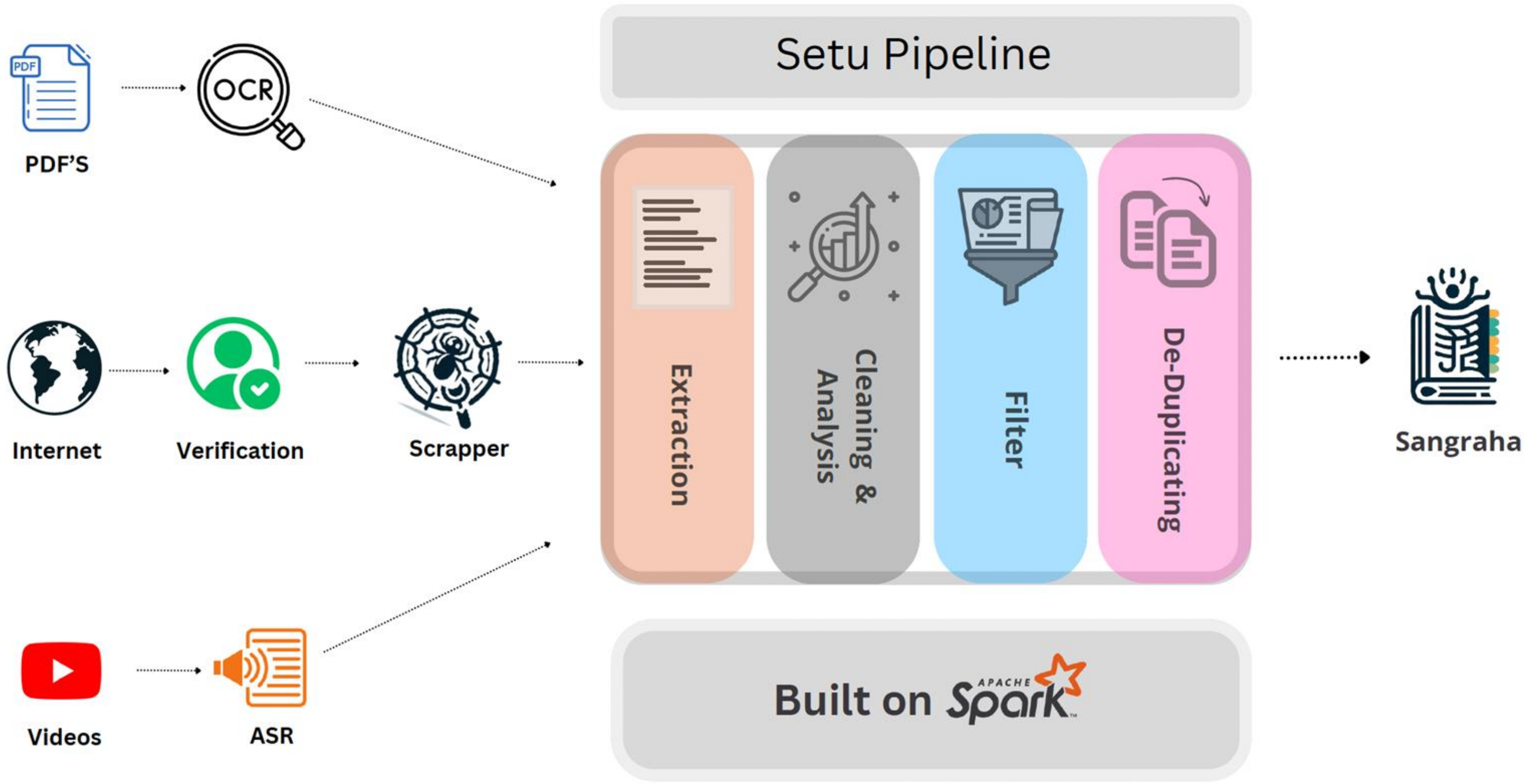
Creating high-quality pre-training data at scale

High –quality sources

Translations and Transliterations to enable cross-lingual transfer

[IndicLLMSuite]

Data Quality is of utmost importance

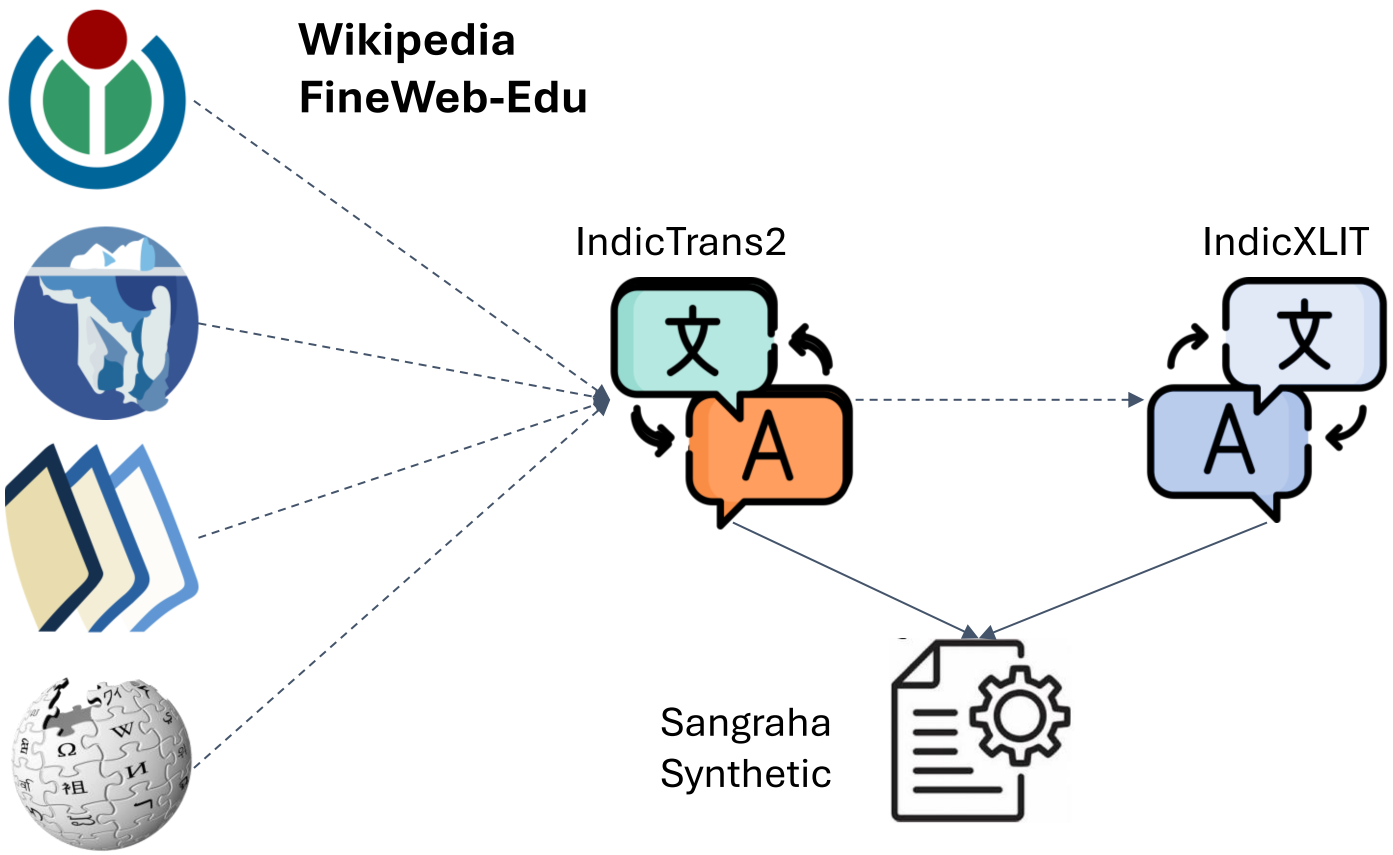


https://github.com/AI4Bharat/setu

Synthetizing Multilingual Data

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

Huge disparity in digital knowledge between English and any other language



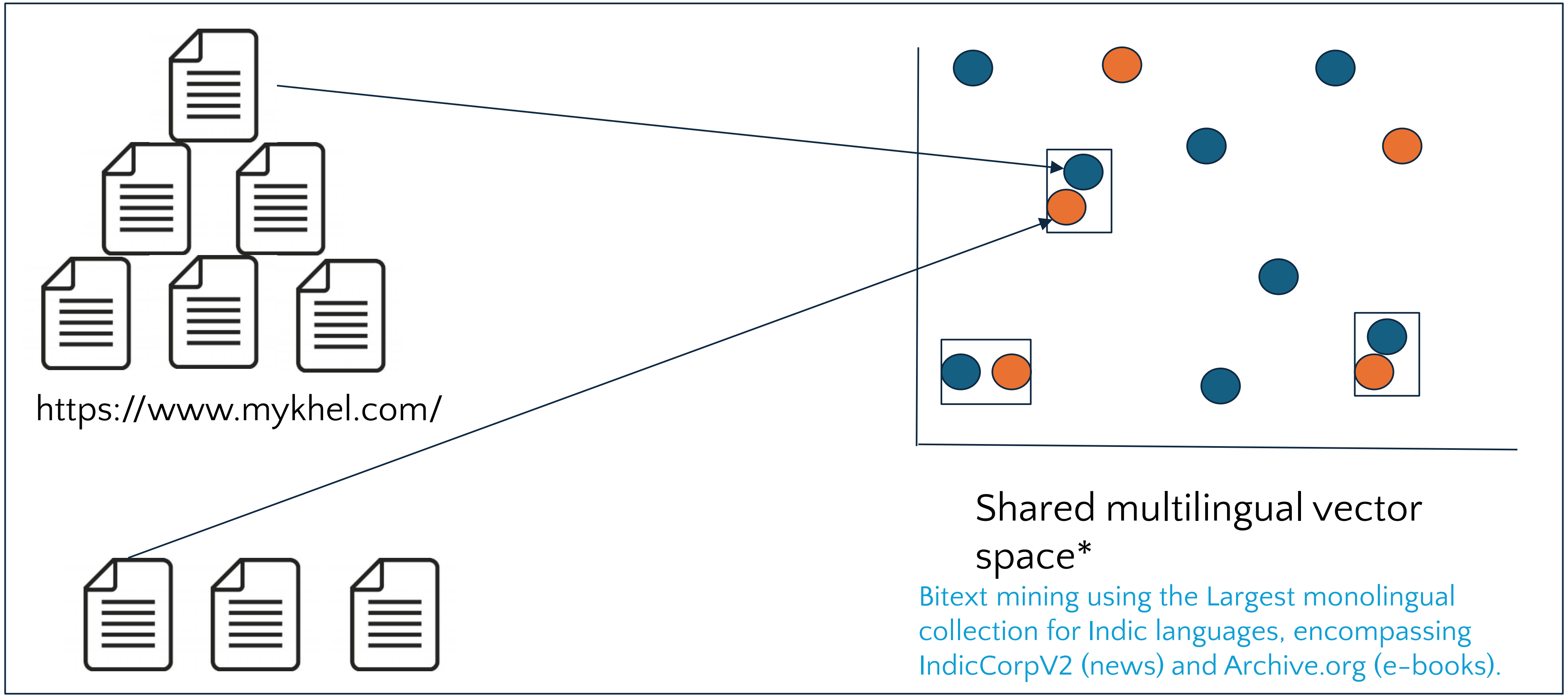
Quick Alternatives →

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

Meet Doshi, et al. "Do Not Worry if You Do Not Have Data: Building Pretrained Language Models Using Translationese." EMNLP 2024.
Jaavid Aktar Husain, et al. "RomanSetu: Efficiently unlocking multilingual capabilities of Large Language Models models via Romanization." ACL 2024.

We also need task-specific dataset for tasks like machine translation, transliteration, ASR, instruction tuning, preference alignment, etc.

Mining Data for training Translation Systems



https://malayalam.mykhel.com/

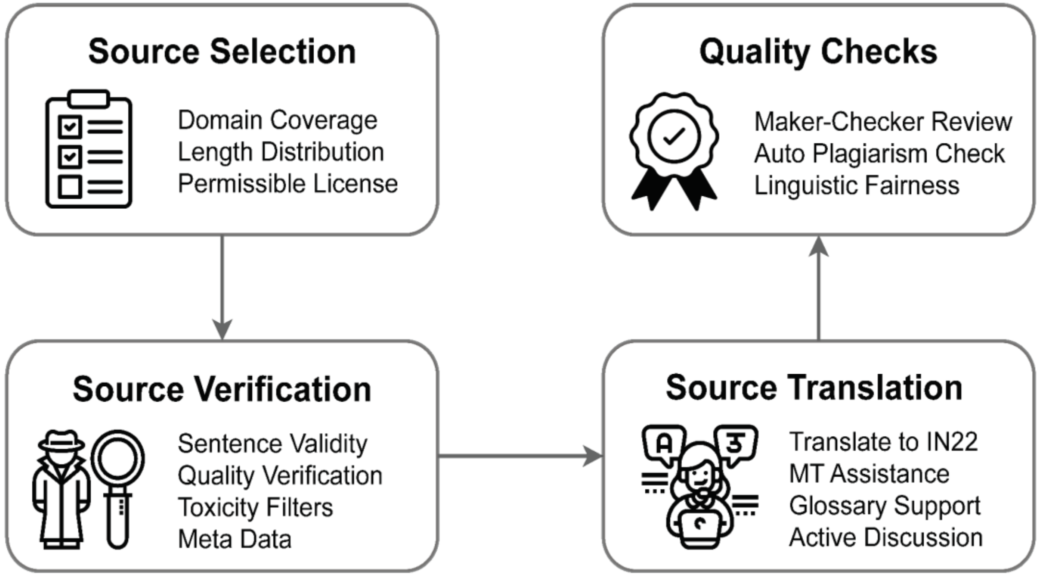
*Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. Language-agnostic BERT sentence embedding, ACL 2022.

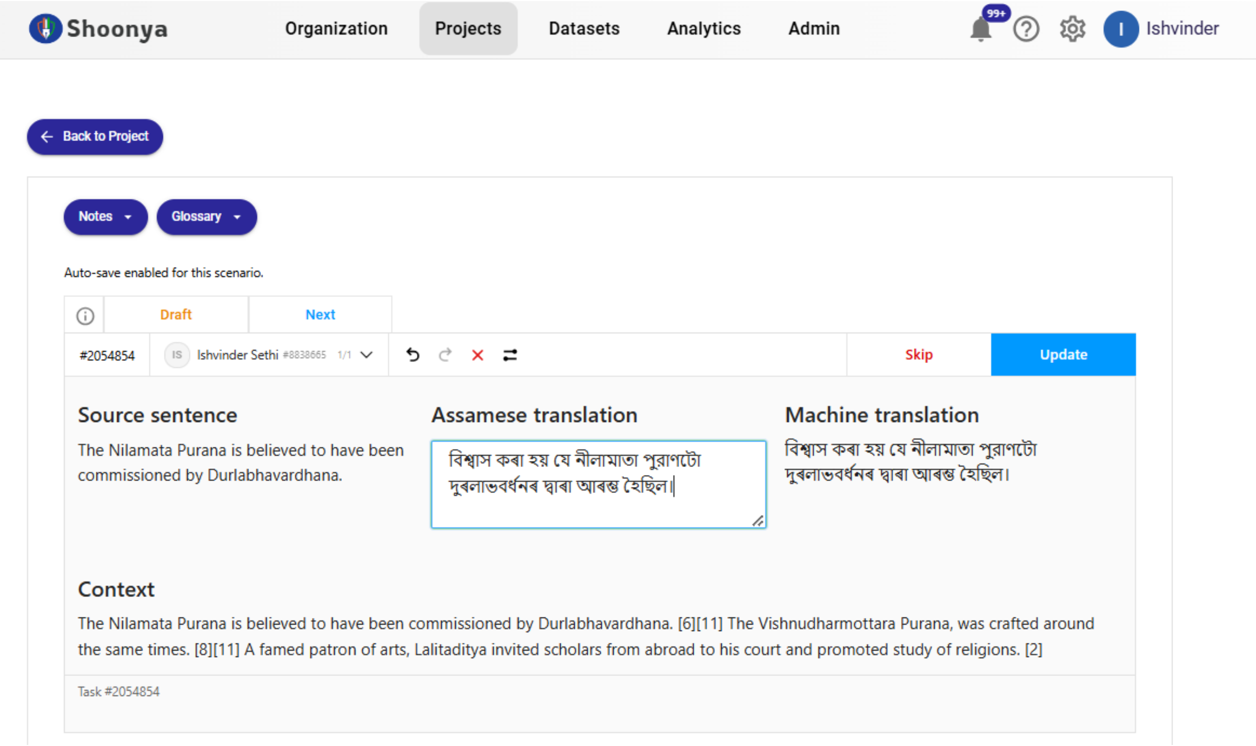
BPCC Corpus: Mined 232 M sentence pairs across 12 Indian languages

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



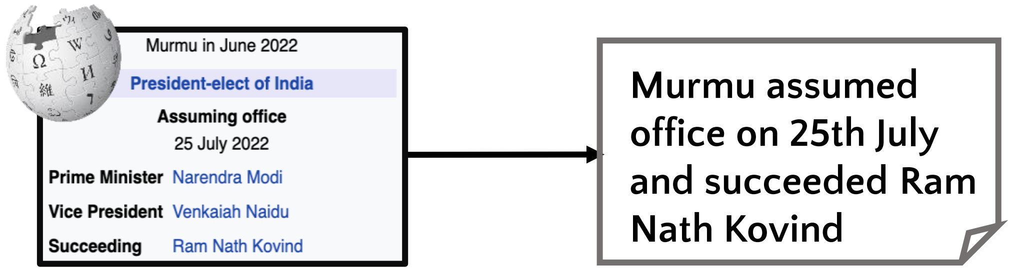


- Need processes in place to ensure high quality
- Provide tools to make translators productive

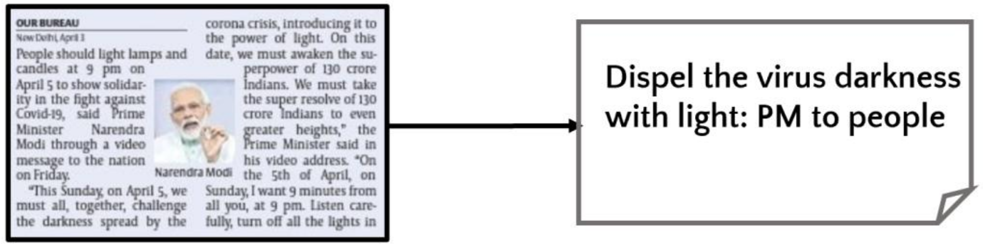
BPCC Corpus: Created 800 K high-quality sentence pairs across 22 Indian languages

Creativity is the limit for mining data of different kinds!

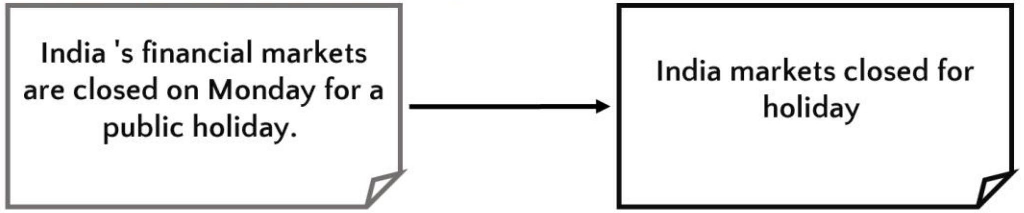
BIOGRAPHY GENERATION



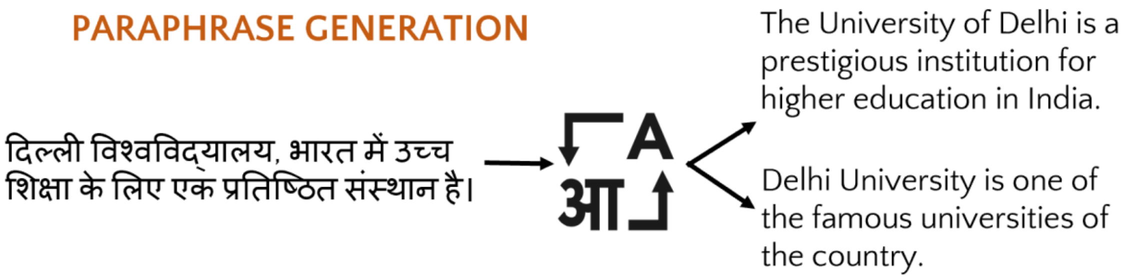
HEADLINE GENERATION



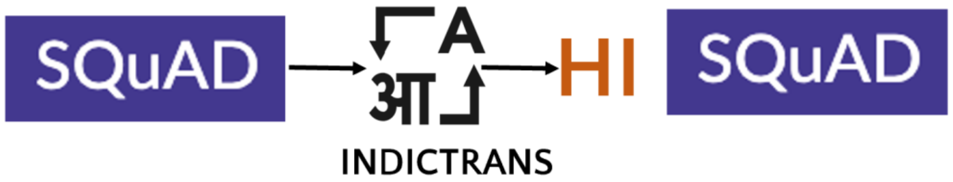
SENTENCE SUMMARIZATION



PARAPHRASE GENERATION



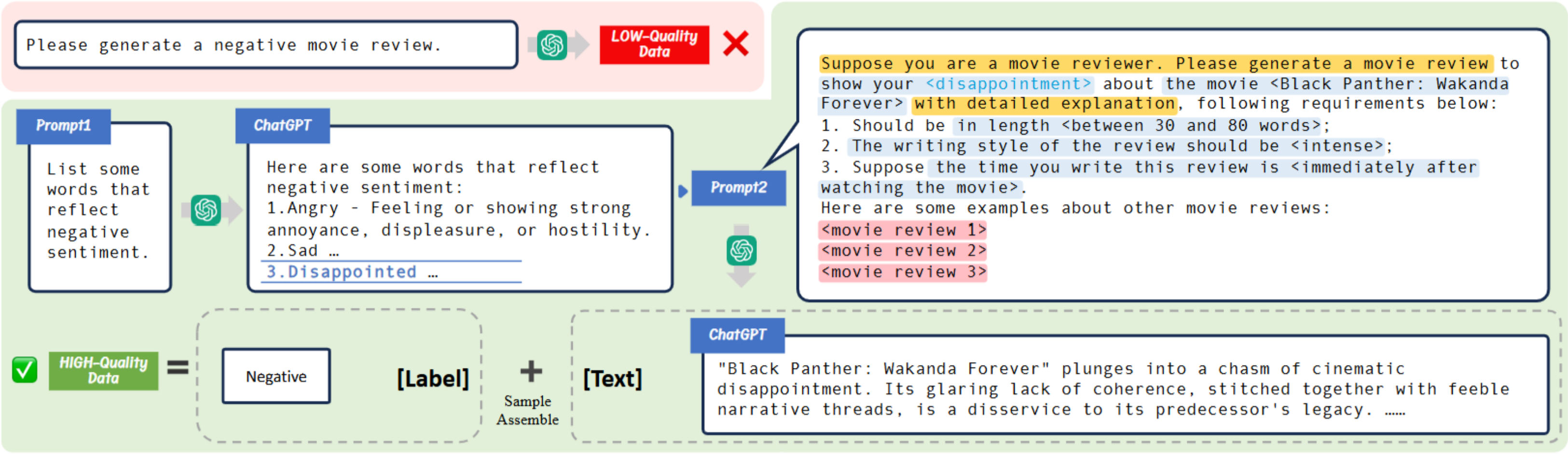
QUESTION GENERATION



Kumar, Aman, et al. "IndicNLG benchmark: Multilingual datasets for diverse NLG tasks in Indic languages." *EMNLP* (2022).

LLMs for Data generation

LLMs have become commonplace for data generation!



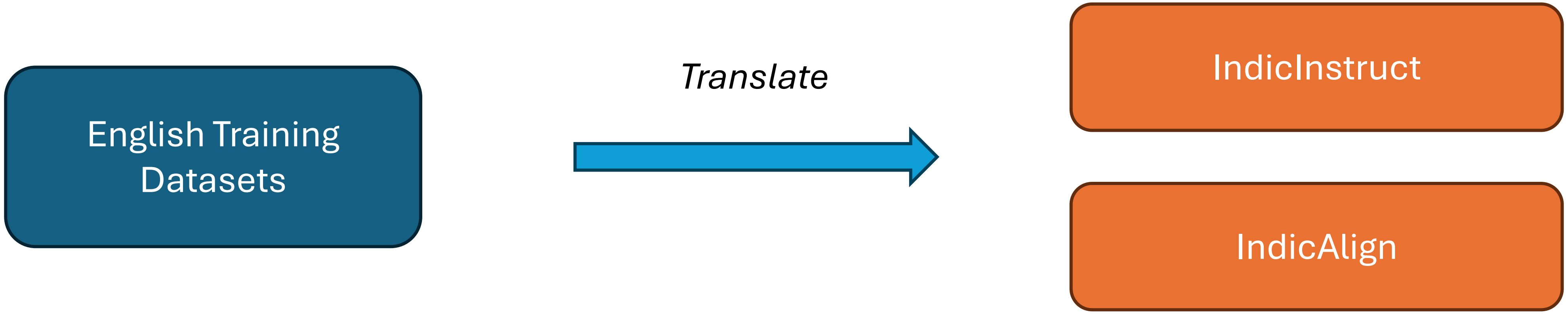
Human creation of data can be laborious, expensive, requires large-scale co-ordination

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

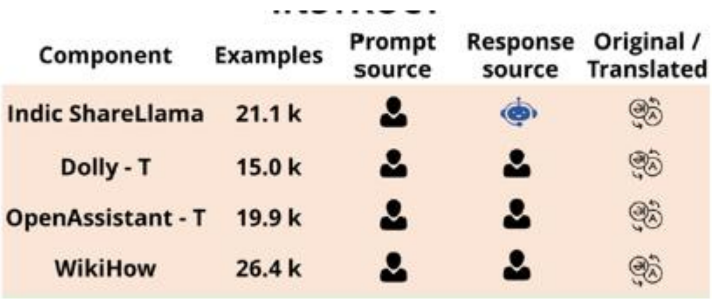
LLMs might not be capable of generating high quality data for non-English languages!

Long, Lin, et al. "On LLMs-driven synthetic data generation, curation, and evaluation: A survey." *arXiv preprint arXiv:2406.15126* (2024).

Machine Translation as an enabler to scaling



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



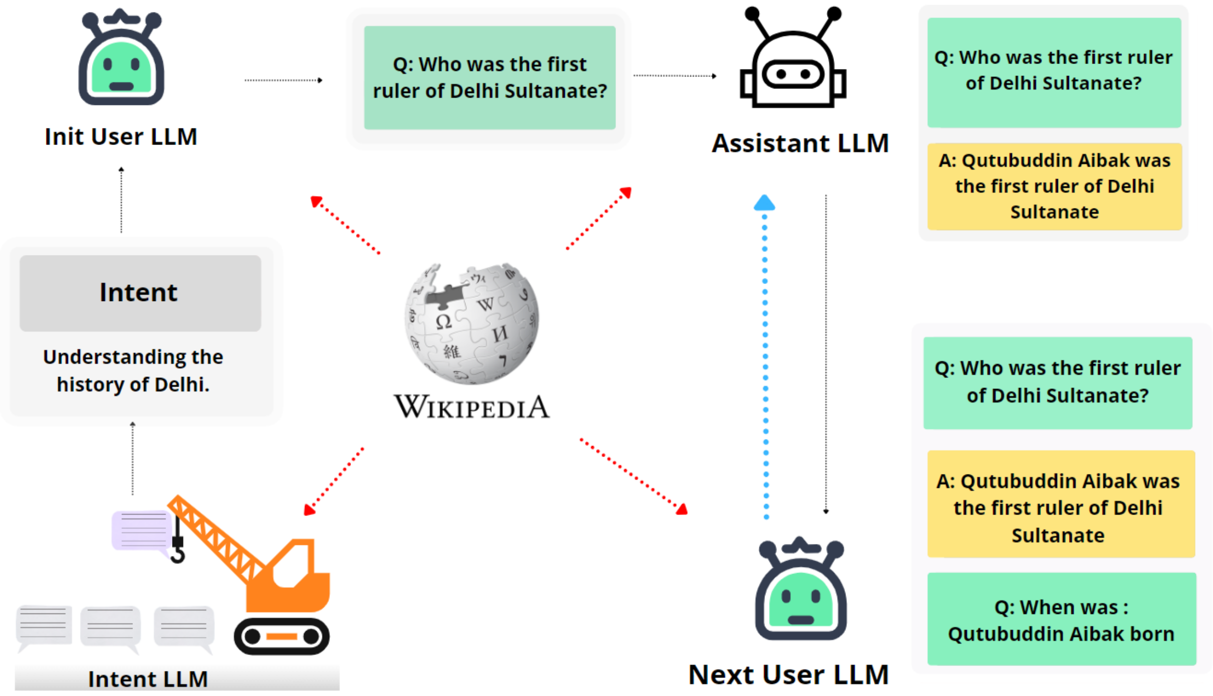
Component	Examples	Prompt source	Response source	Original / Translated
Indic ShareLlama	21.1 k			
Dolly - T	15.0 k			
OpenAssistant - T	19.9 k			
WikiHow	26.4 k			

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

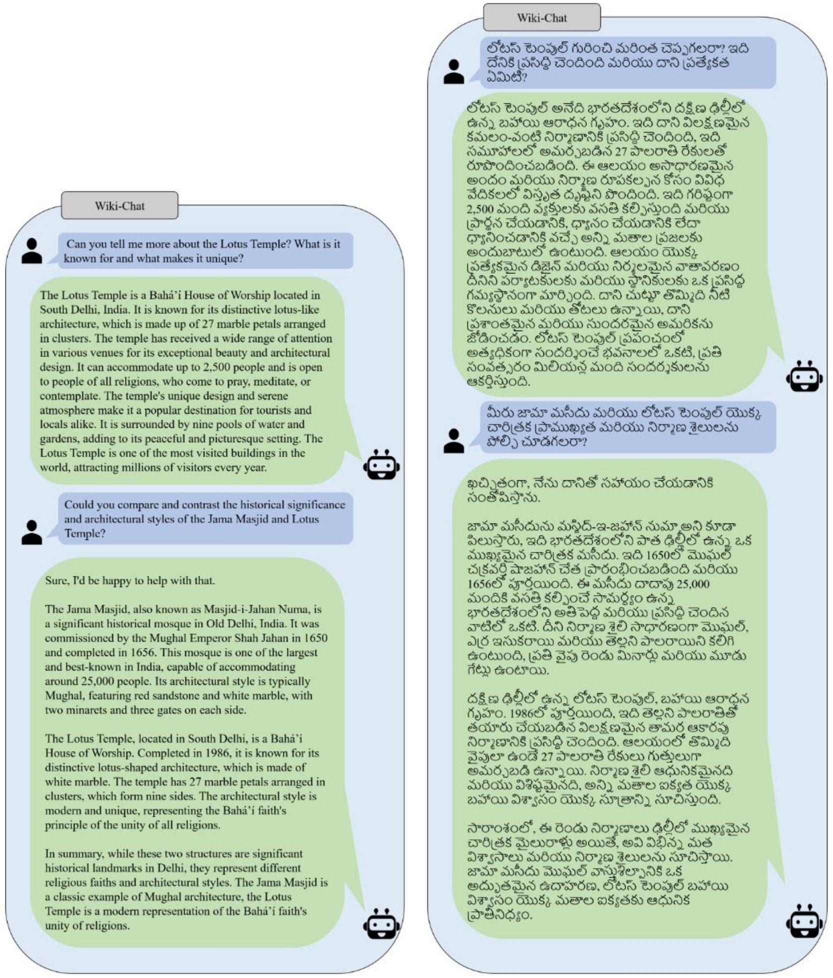
Creativity is the limit

Generating Culturally Relevant IFT Data

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



Translate the conversations from English

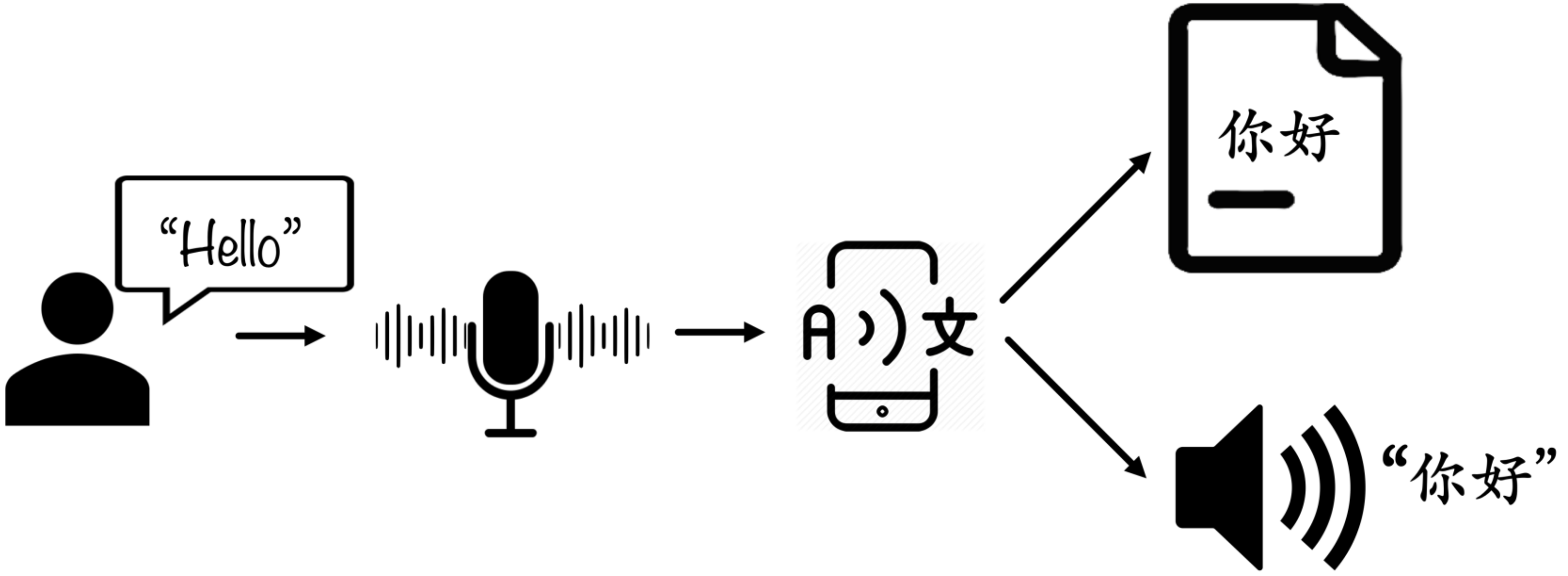


(a) English (b) Telugu

(IndicLLMSuite)

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*

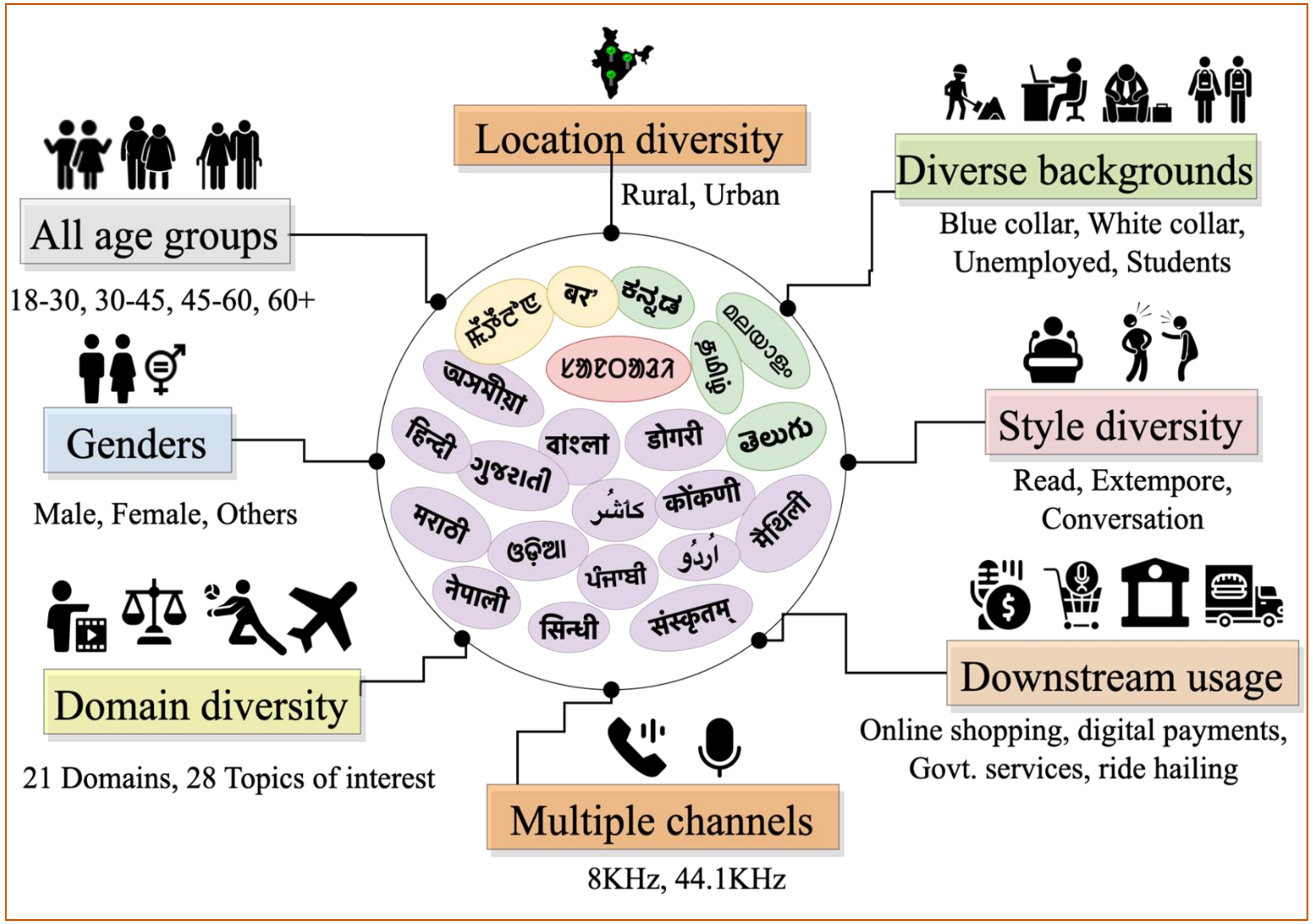
***Speeches/Podcasts** like TED, VaaniPedia, Mann ki Baat*

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

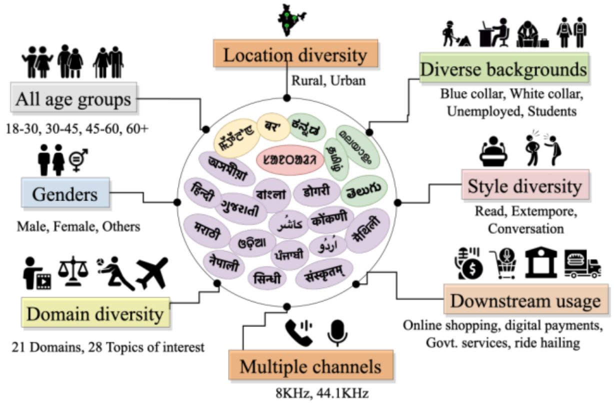


- District wise collection
- Ensuring inclusivity
- Ensuring diversity

Three Key Contributions!

Data	Model	Starter kit

Data



IndicVoices

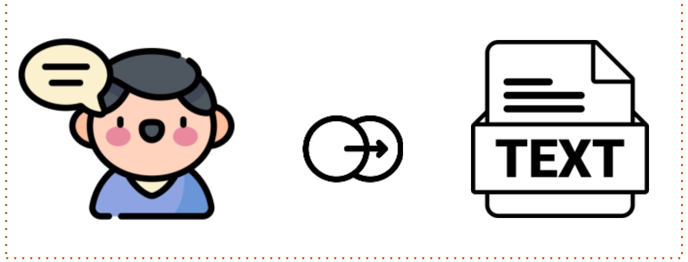
*(Goal: **17000** hours)*

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

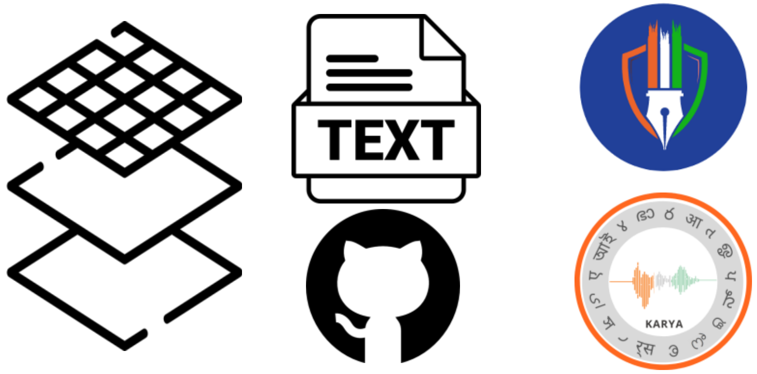
Model

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

Opportunities

- Building preference alignment datasets for Indian languages considering cultural nuances, viewpoints, local use-cases, etc.
- Building datasets for specialized areas important in the Indian context with domain experts.
- Share specialized datasets, unlock private data sources to drive research and innovation

Multi-linguality and Indic language viewpoint

Are Indian languages related?

Can similarity between languages be used to build better language technology?

There is unity in Indian languages

Related Languages

Related by Genealogy

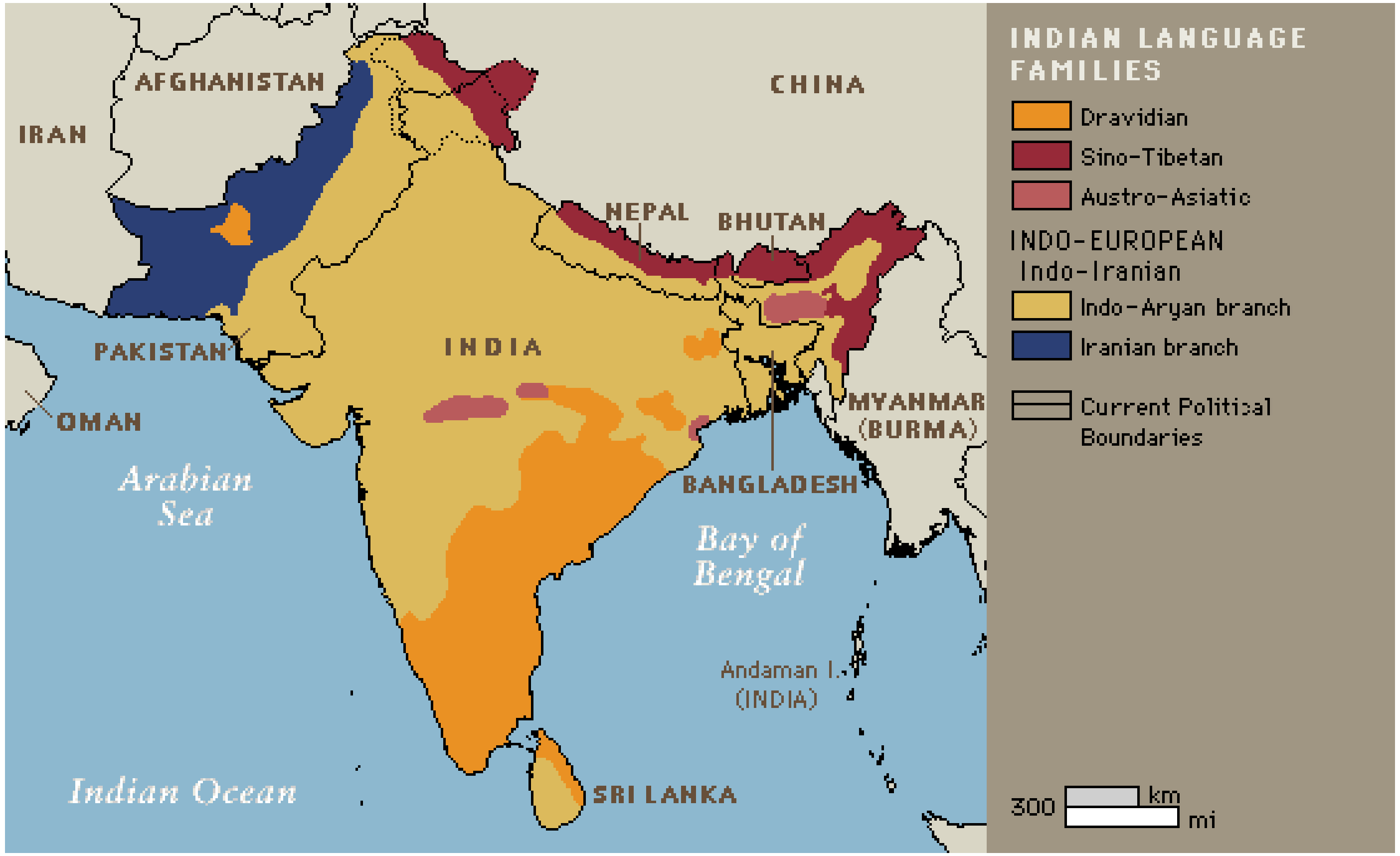
Related by Contact

<u>Language Families</u>
Dravidian, Indo-European, Turkic

<u>Linguistic Areas</u>
Indian Subcontinent, Standard Average European

Related languages may not belong to the same language family!

Language Families in India



<u>4 major language families</u>

<u>Indo-Aryan</u>: *North India and Sri Lanka (branch of Indo-European)*

<u>Dravidian</u>: *South India & pockets in the North*

<u>Tibeto-Burman</u>: *North-East and along the Himalayan ranges*

<u>Austro-Asiatic</u>: *pockets in Central India, North-East, Nicobar Islands*

\+

Andamanese family

Unknown language of the Sentinelese

Cognates & Borrowed words in Indian Languages

Indo-Aryan

English	Vedic Sanskrit	Hindi	Punjabi	Gujarati	Marathi	Odia	Bengali
bread	Rotika	chapātī, roṭī	roṭi	paũ, roṭlā	chapāti, poli, bhākarī	pauruṭi	(pau-)ruṭi
fish	Matsya	Machhlī	machhī	māchhli	māsa	mācha	machh
hunger	bubuksha, kshudhā	Bhūkh	pukh	bhukh	bhūkh	bhoka	khide

Dravidian

English	**Tamil**	**Malayalam**	**Kannada**	**Telugu**
fruit	pazham , kanni	pazha.n , phala.n	haNNu , phala	pa.nDu , phala.n
ten	pattu	patt,dasha.m,dashaka.m	hattu	padi

Indo-Aryan words in Dravidian languages

Sanskrit word	**Language**	**Loanword**	**English**
cakram	Tamil	cakkaram	wheel
matsyah	Telugu	matsyalu	fish
ashvah	Kannada	ashva	horse
jalam	Malayalam	jala.m	water

Other borrowings like echo words, retroflex sounds in other direction. (Subbarao, 2012)

Source: Wikipedia and IndoWordNet

Key Similarities between related languages

भारताच्या स्वातंत्र्यदिनानिमित्त अमेरिकेतील लॉस एन्जल्स शहरात कार्यक्रम आयोजित करण्यात आला
bhAratAcyA svAta.ntryadinAnimitta ameriketIla lOsa enjalsa shaharAta kAryakrama Ayojita karaNyAta AlA

Marathi

भारता च्या स्वातंत्र्य दिना निमित्त अमेरिके तील लॉस एन्जल्स शहरा त कार्यक्रम आयोजित करण्यात आला
bhAratA cyA svAta.ntrya dinA nimitta amerike tIla lOsa enjalsa shaharA ta kAryakrama Ayojita karaNyAta AlA

Marathi segmented

भारत के स्वतंत्रता दिवस के अवसर पर अमरीका के लॉस एन्जल्स शहर में कार्यक्रम आयोजित किया गया
bhArata ke svata.ntratA divasa ke avasara para amarIkA ke losa enjalsa shahara me.n kAryakrama Ayojita kiyA gayA

Hindi

Lexical: share significant vocabulary (cognates & loanwords)

Morphological: correspondence between suffixes/post-positions

Syntactic: share the same basic word order

Similarity of Indian Scripts

- **Largely overlapping character set**, but the visual rendering differs
- Traditional ordering of characters is same (*varnamala*)
- Dependent (*maatras*) and Independent vowels

***Abugida* scripts:**

- primary consonants with secondary vowels diacritics (*maatras*)
- rarely found outside of the Brahmi family

- Consonant clusters (क्क,क्ष)
- Special symbols like:
  - *anusvaara* (nasalization), *visarga* (aspiration)
  - *halanta/pulli* (vowel suppression), nukta (Persian/Arabic sounds)
- Basic Unit is the akshar (a pseudo-syllable)

India as a linguistic area gives us robust reasons for writing a common or core grammar of many of the languages in contact

~ Anvita Abbi

Are Indian languages related?

Can similarity between languages be used to build better language technology?

Script Conversion

- Read any script in any script
- Unicode standard enables consistent script conversion

$$\textit{unicode\_codepoint(char)} - \textit{Unicode\_range\_start}(L_1) + \textit{Unicode\_range\_start}(L_2)$$

केरला

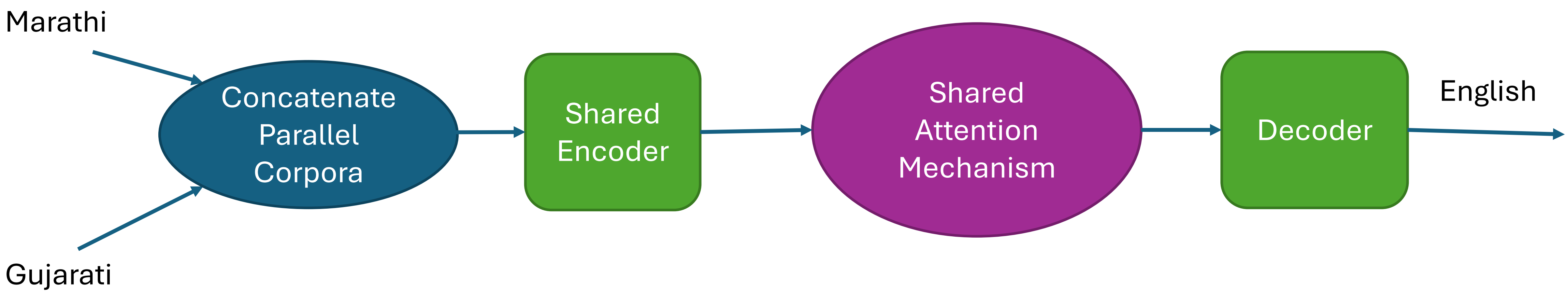
কেরলা

કેરલા

Multilingual Neural Machine Translation

(Zoph et al., 2016; Nguyen et al., 2017; Lee et al., 2017; Dabre et al., 2018)

We want Gujarati → English translation ➔ but little parallel corpus is available
We have lot of Marathi → English parallel corpus



Combine Corpora from different languages

(Nguyen and Chang, 2017)

I am going home	હુ ઘરે જવ છૂ
It rained last week	છેલ્લા આઠવડિયા મા વર્સાદ પાડ્યો

It is cold in Pune	पुण्यात थंड आहे
My home is near the market	माझा घर बाजाराजवळ आहे

Convert Script

Concat Corpora

I am going home	हु घरे जव छू
It rained last week	छेल्ला आठवडिया मा वर्साद पाड्यो
It is cold in Pune	पुण्यात थंड आहे
My home is near the market	माझा घर बाजाराजवळ आहे

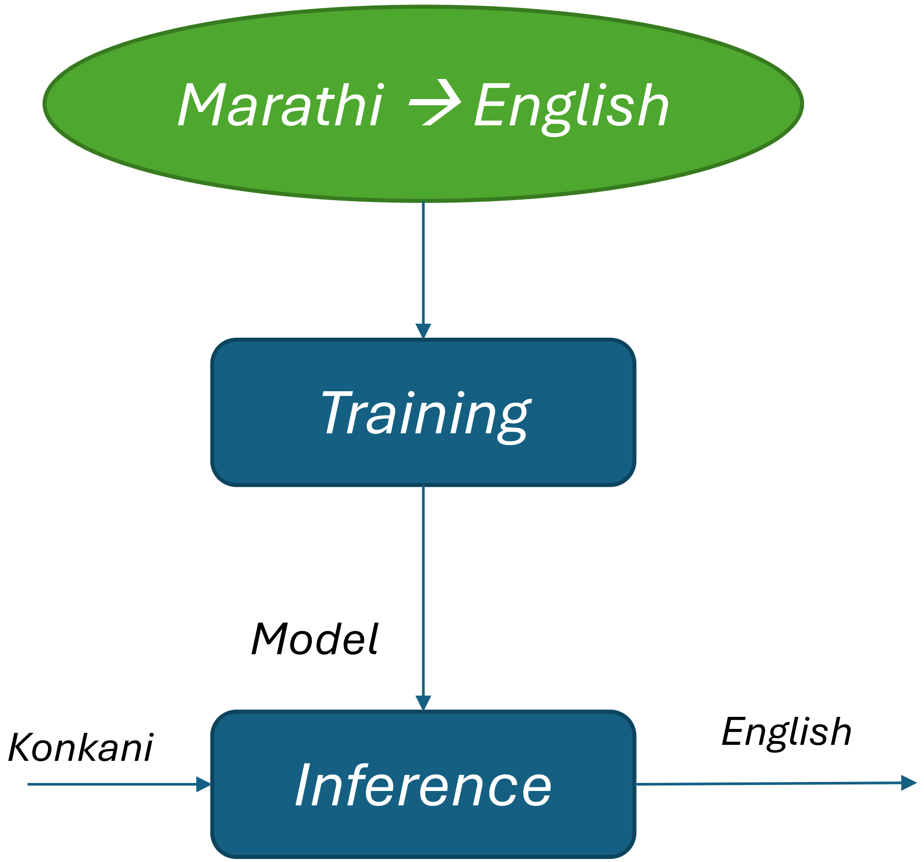
Significant boost in quality for the many languages, particularly the ones which have little data

Better generalization due to exposure to diverse data

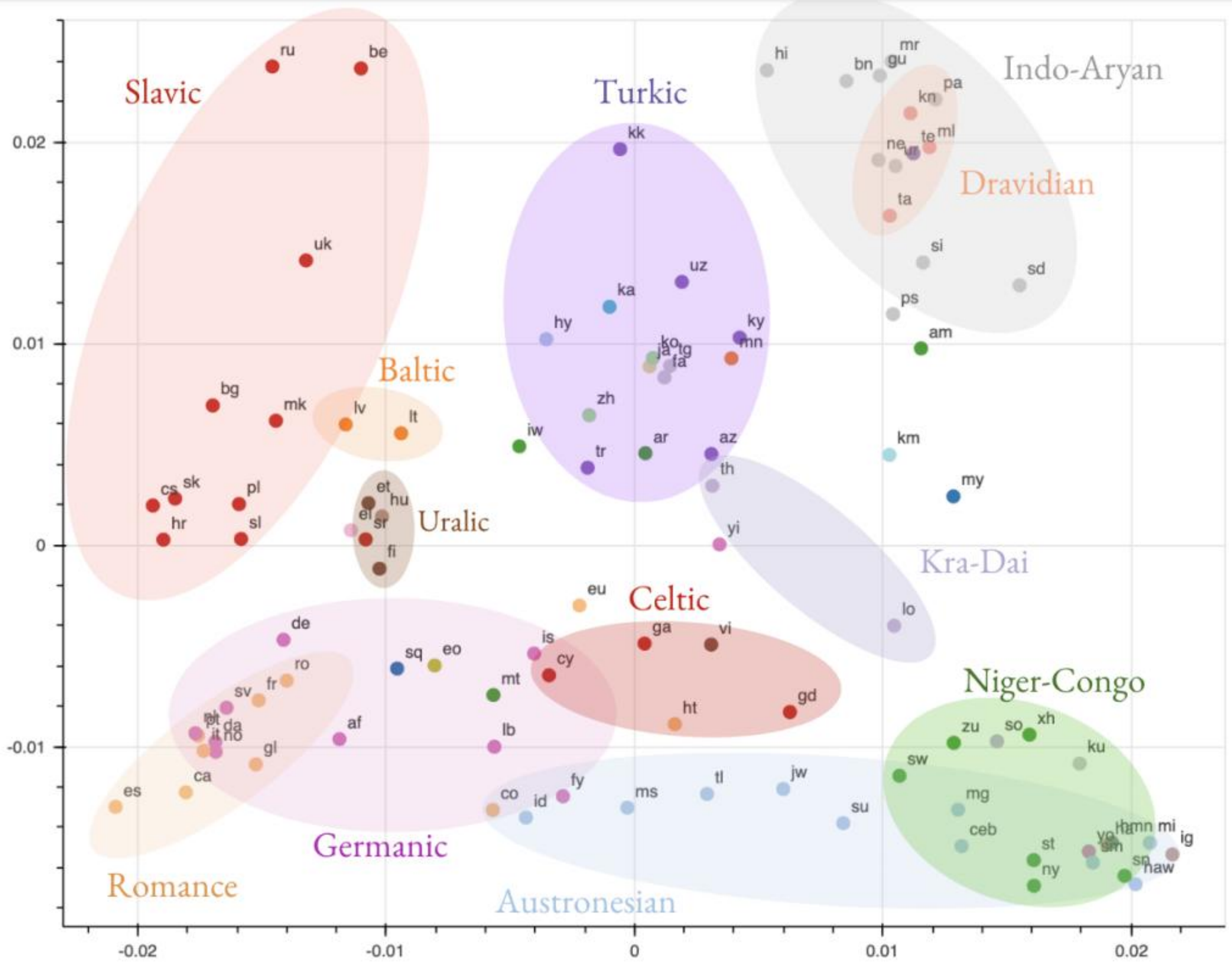
Zeroshot Translation

For very closely related languages,

Systems can perform well with very little or no data



Transfer Learning works best for related languages



(Kudungta et al, 2019)

The central goal in multilingual NLP is to align language representations to be agnostic to language

Maximize transfer learning benefits

Summary

- **Deep Learning has revolutionized** multilingual representation learning
- **Opened up many possibilities**
  - Multilingual training, Zero-shot performance, Compact models
  - Support low-resource languages and domains via transfer Learning

Opportunities

- Multilingual transfer from English
- Multilingual knowledge transfer
- Transfer in preference alignment models
- Multilingual reasoning

Time to look at efficiency at all stages of LLM development

LLM Training

- Trillions of words
- Hundreds of thousands of GPUs
- Lot of manpower, expertise for high quality data collection
- Lot of experimentation

Challenges →

- Sustainable research and development cycle
- Energy Consumption
- Rapid Adaptation

Fairly unexplored in the Indian context, lots of opportunities

Data-efficient learning

Efficient model architectures & learning

Systems optimization

Efficient inference

Model Adaptation

AI aided Data Annotation

Support new languages, modalities and domains

Glimpses of Indic Language Tech Efforts

A snapshot of AI4Bharat Efforts

NLP Infrastructure: Raw corpora

IndicCorp

Large Monolingual corpora

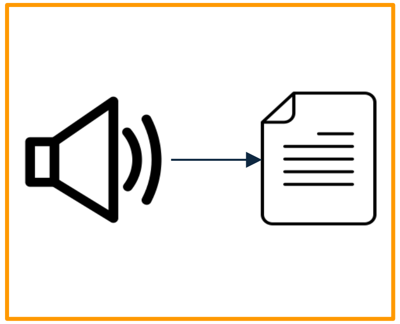
20B tokens, 22 languages

Sangraha

Large Document

Monolingual corpora

100B tokens, 22 languages



MahaDhwani

Raw speech corpora

(279k hours, 22

languages)

NLP Infrastructure: language models

IndicFT

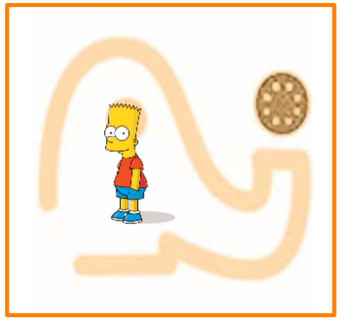
(word embeddings)

IndicWav2Vec

(Pre-trained speech model)

IndicBERT

(encoder LM)



IndicBART

(seq2seq LM)

Airavata

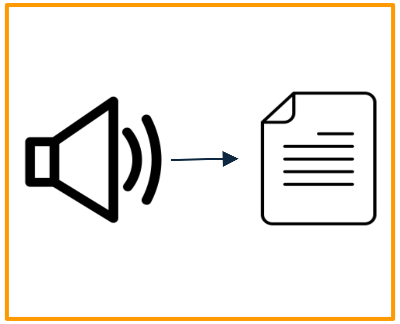
(Finetuned LLM)

Compact pre-trained models for NLU & NLG

Data for various foundational tasks

BPCC

Parallel corpus, translation models between English & 22 Indic languages

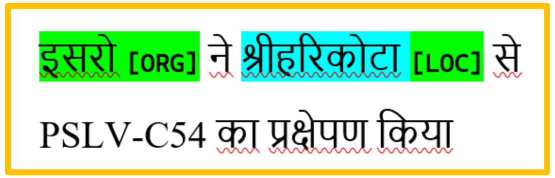


Shrutilipi, IndicVoices & KathBath

ASR datasets for 22 Indian languages

Aksharantar

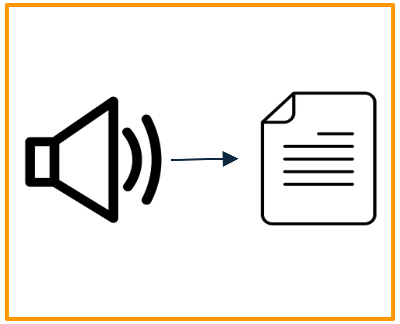
Transliteration datasets for 20 Indic languages



Naamapadam

Datasets and models for Named Entity Recognition in 11 Indian languages

Models for fundamental tasks



इसरो [ORG] ने श्रीहरिकोटा [LOC] से PSLV-C54 का प्रक्षेपण किया

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*

Summary and Outlook

- Language Technology is at **center of modern AI advances**
- Indian languages present both an **opportunity and challenge**
- **Rapid advances** in recent years in ILT, but we have **moving targets**
- **Collaboration and open R&D** to stay ahead of the curve
- **Efficiency in model development** important for sustainable progress in the field
- **The Human Factor**: Tech skilling as well building and utilizing domain expertise for creating datasets are important

Thanks

anoop.kunchukuttan@gmail.com

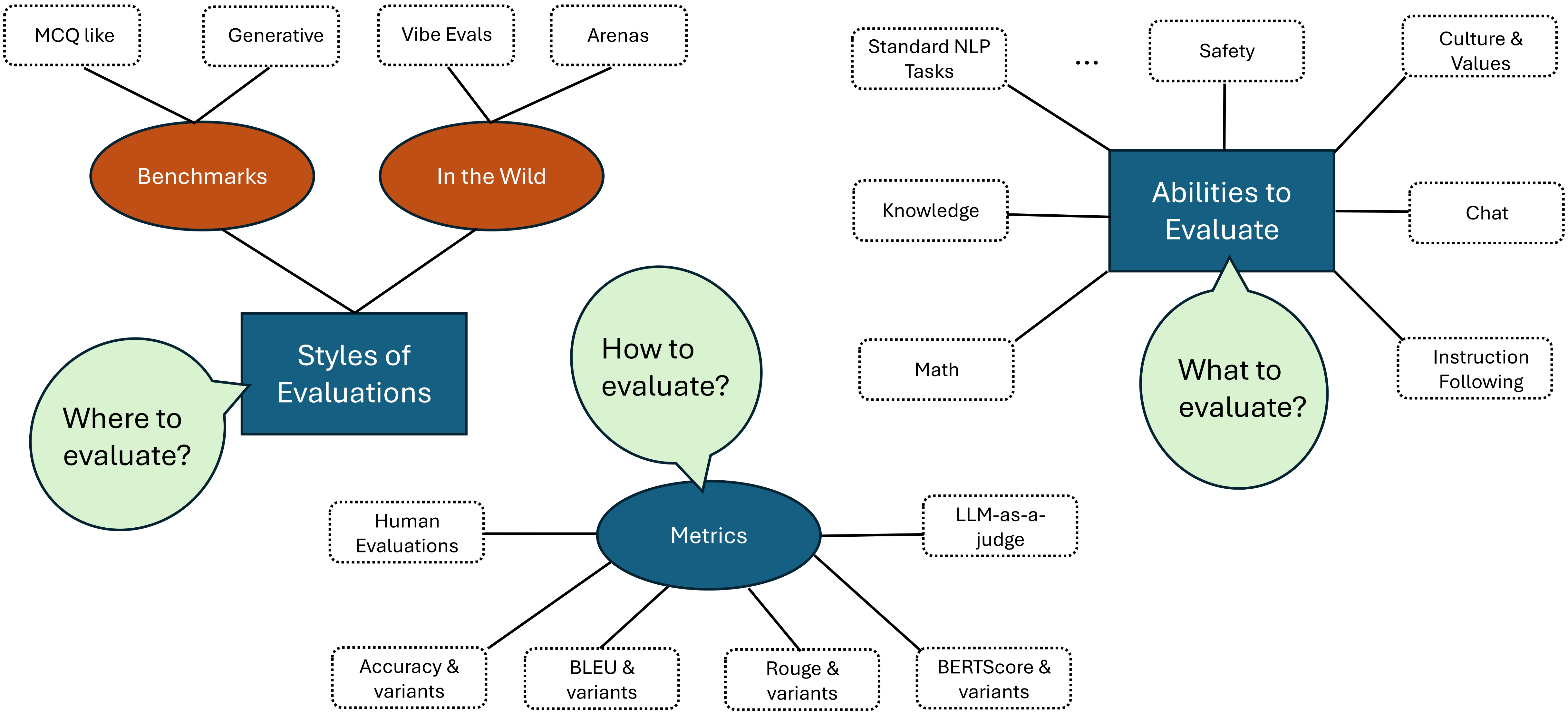
http://anoopkunchukuttan.github.io

http://huggingface.co/AI4Bharat

Acknowledgments: All my collaborators, colleagues and students at AI4Bharat and Microsoft

Evaluation

Broad Taxonomy of Evaluations



Current State of Multilingual Benchmarks

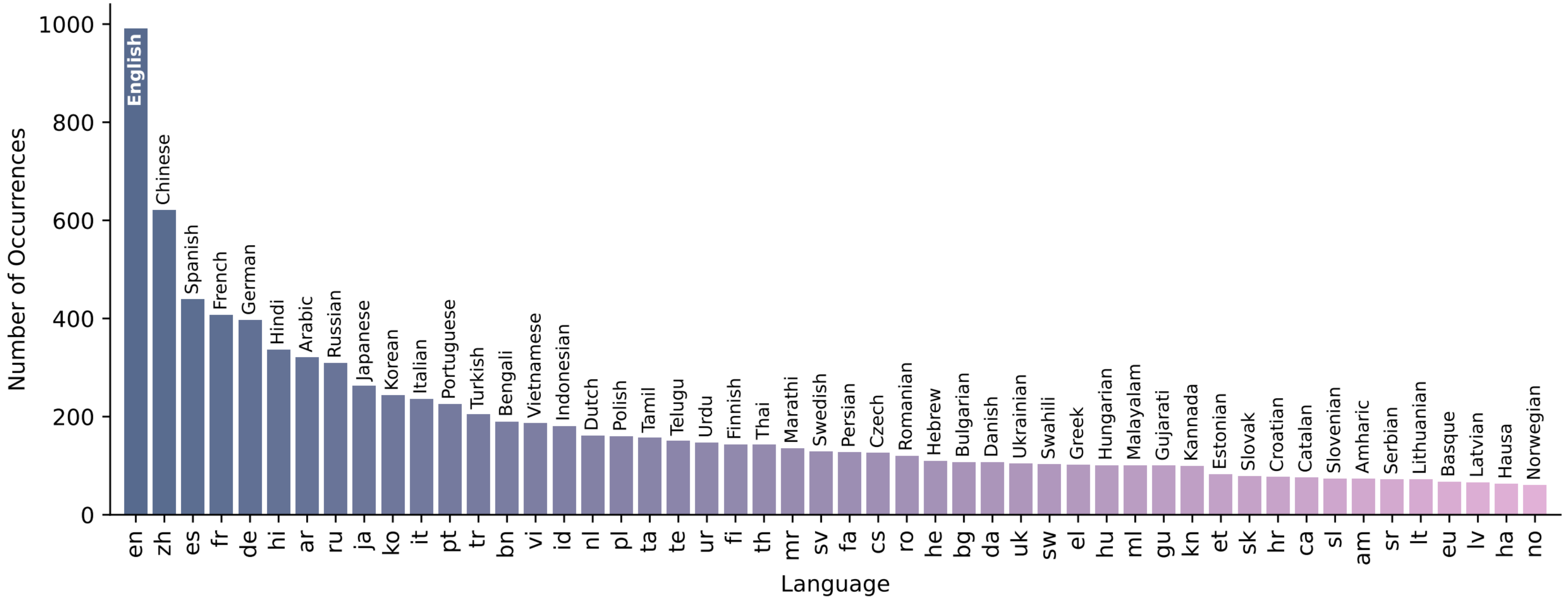
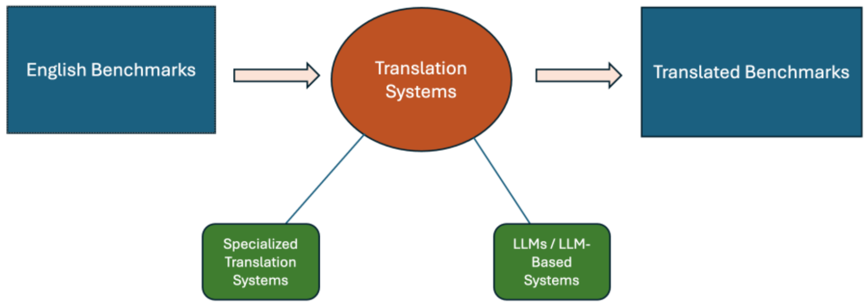


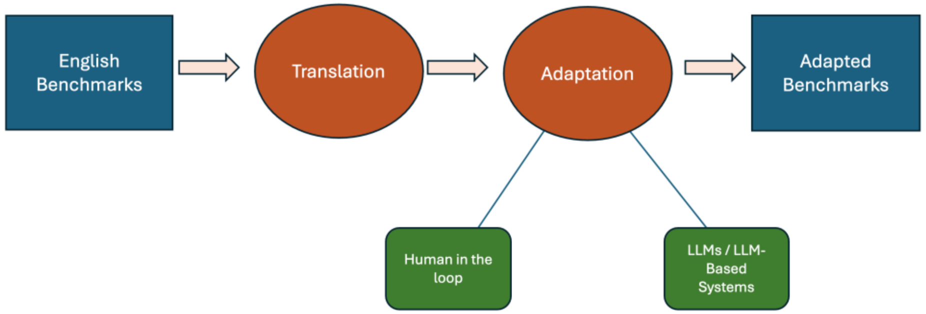
Figure 2 | Distribution of the top 50 languages in our multilingual benchmark collection. Although English is deliberately excluded from the collection, it still appears as the most frequent language in the collection. This distribution illustrates the current imbalance in multilingual evaluation benchmarks.

[BitterLesson]

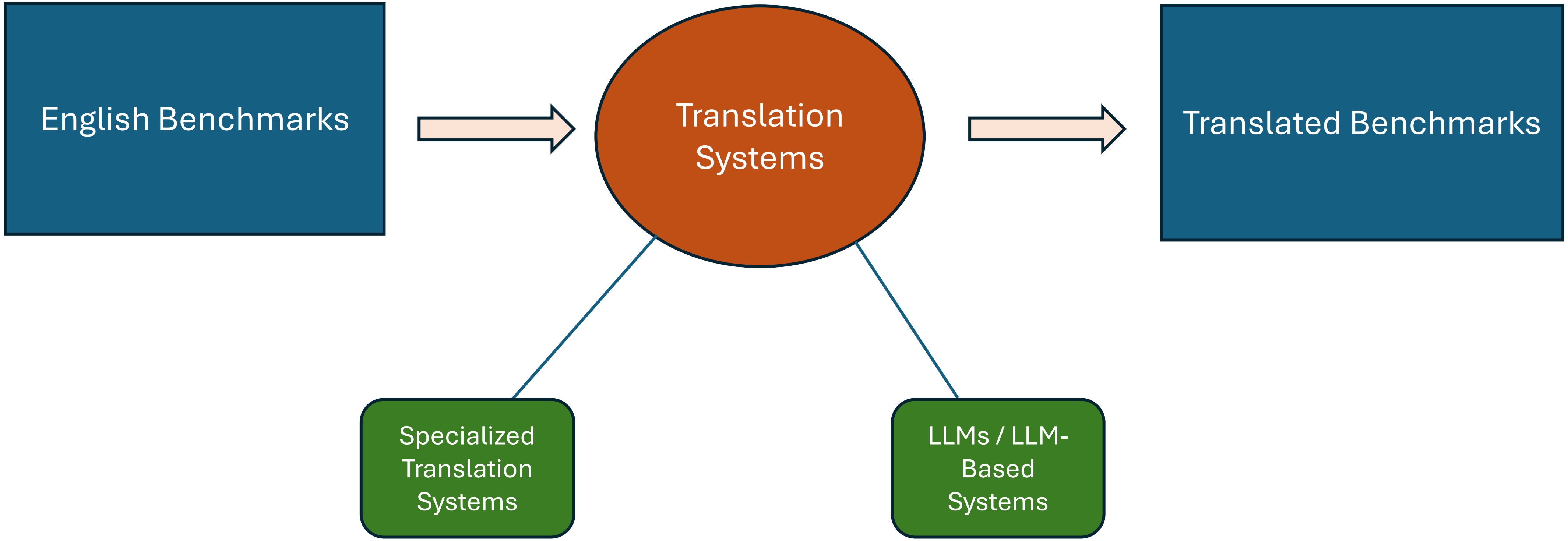
Strategies for Creating Benchmarks

- Machine Translation
- Adaptation
- Create from Scratch





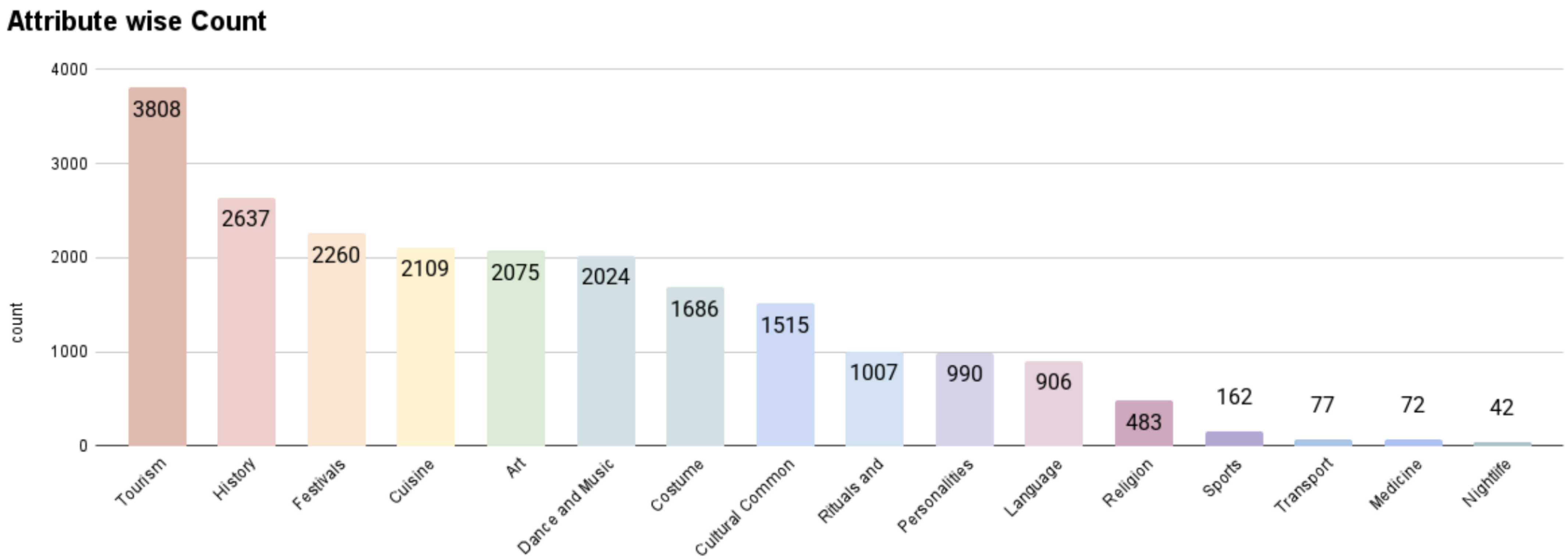
Translation



Very high quality translations required

Human-in-the-loop approaches and sophisticated translation pipelines needed

Cultural Evaluation



[SANSKRITI]

Summary & Recommendations

- **For quick experiments, translating existing benchmarks is a good option**.
- **Ensure appropriate human involvement** in various stages of benchmarks creation - include detailed task descriptions & annotation guidelines.
- **Exercise caution while using LLM-as-a-judge approach**. Thoroughly test for human correlations for your tasks & languages.
- Ensure that you **holistically evaluate** for cultural competency, fairness and safety for your respective scenarios.