NLP At Scale for Indian Languages

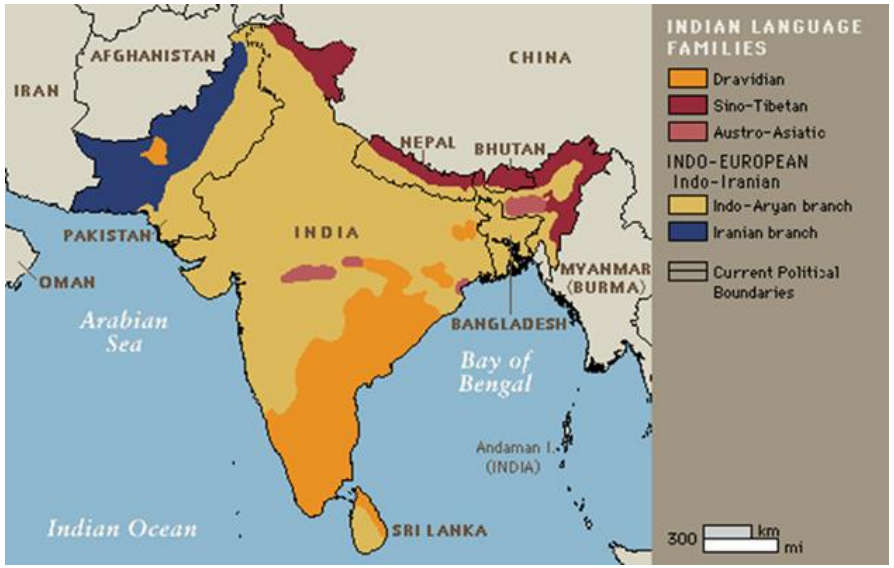
Anoop Kunchukuttan

Microsoft Translator

AI4Bharat

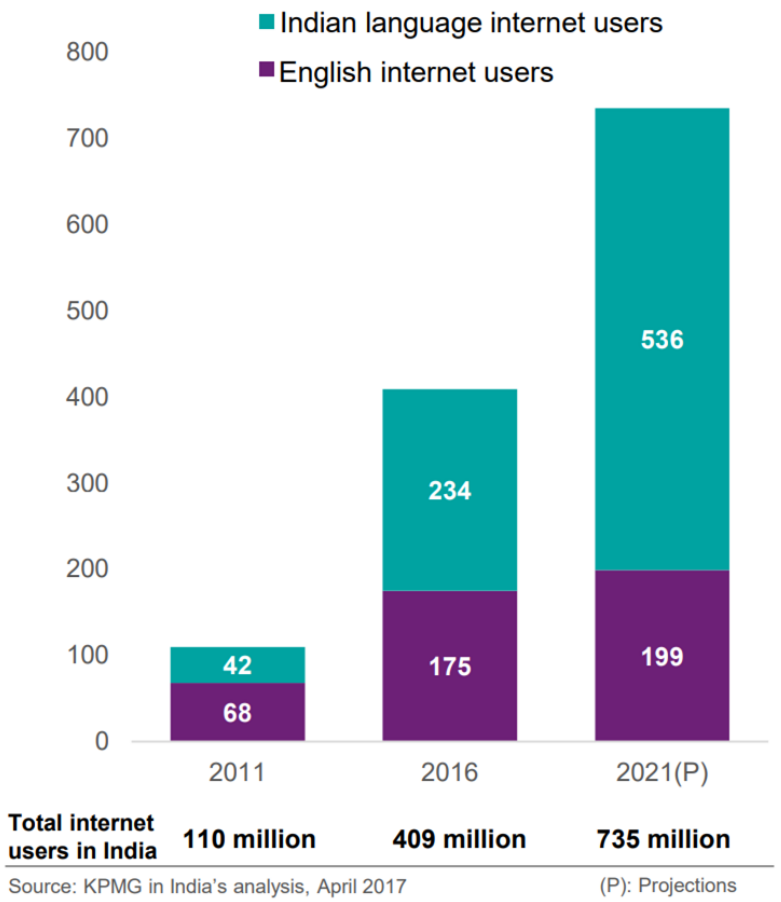
LTRC Silver Jubilee, IIIT Hyderabad, Dec 6-7 2024

Usage and Diversity of Indian Languages



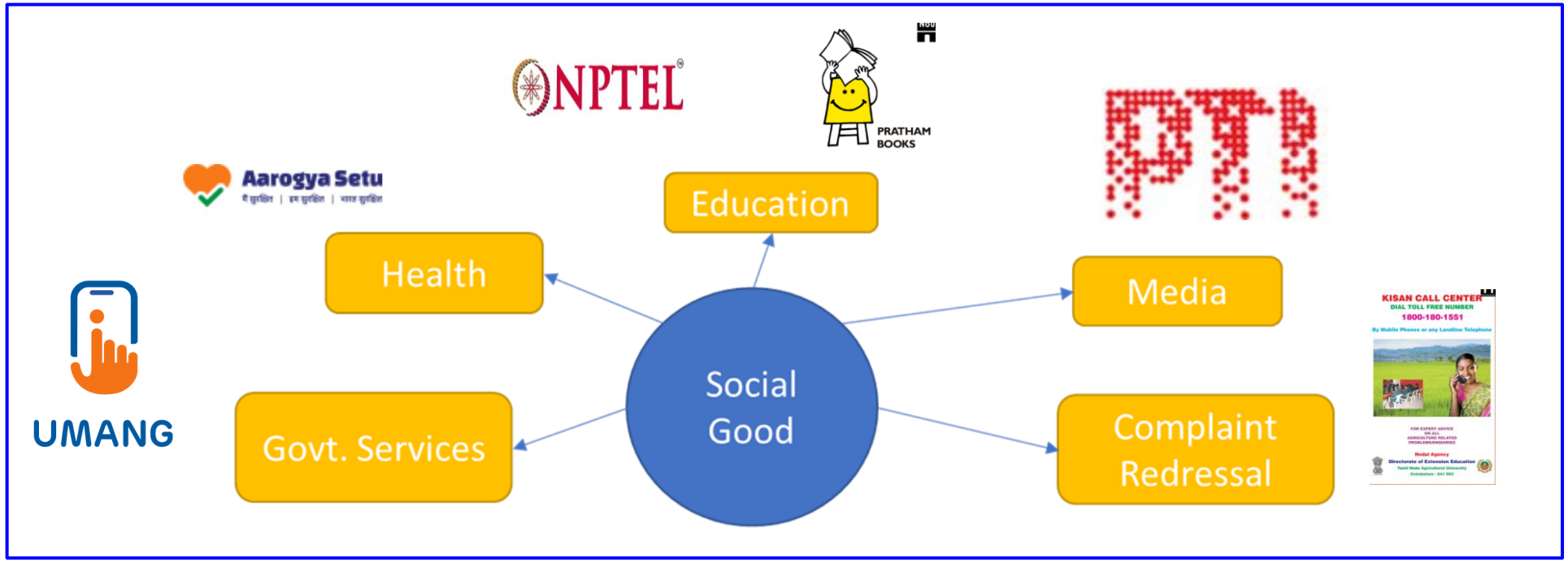
- *4 major language families*
- *22 scheduled languages*
- *125 million English speakers*
- *8 languages in the world's top 20 languages*
- *30 languages with more than 1 million speakers*

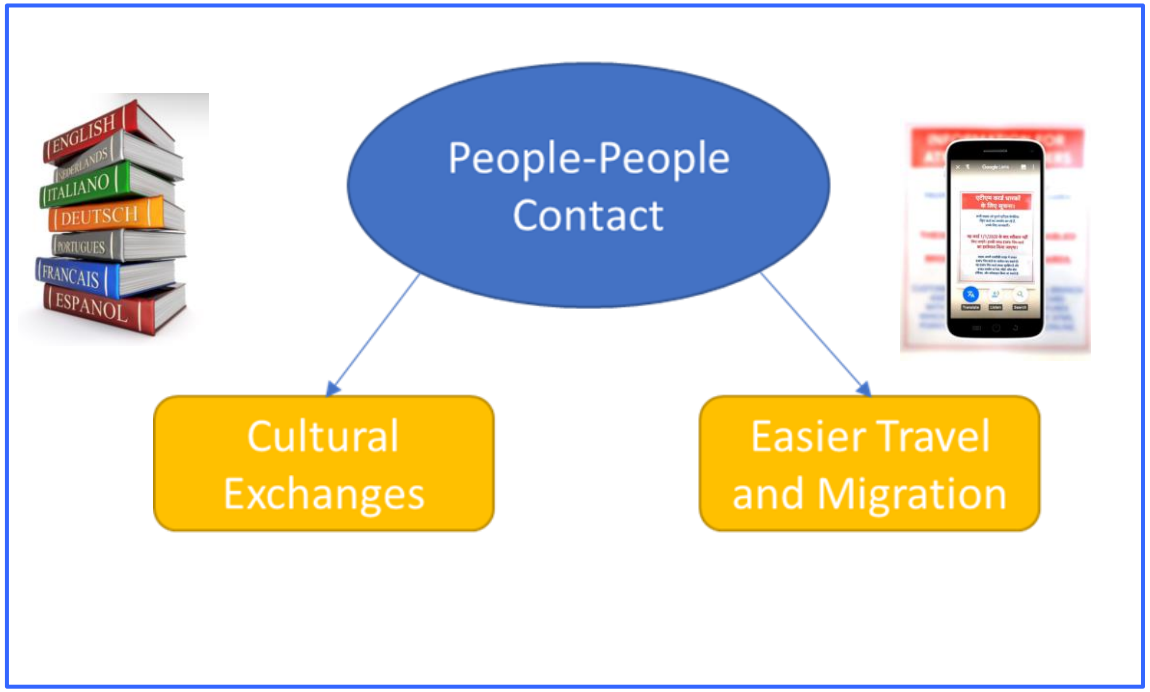
Sources: Wikipedia, Census of India 2011

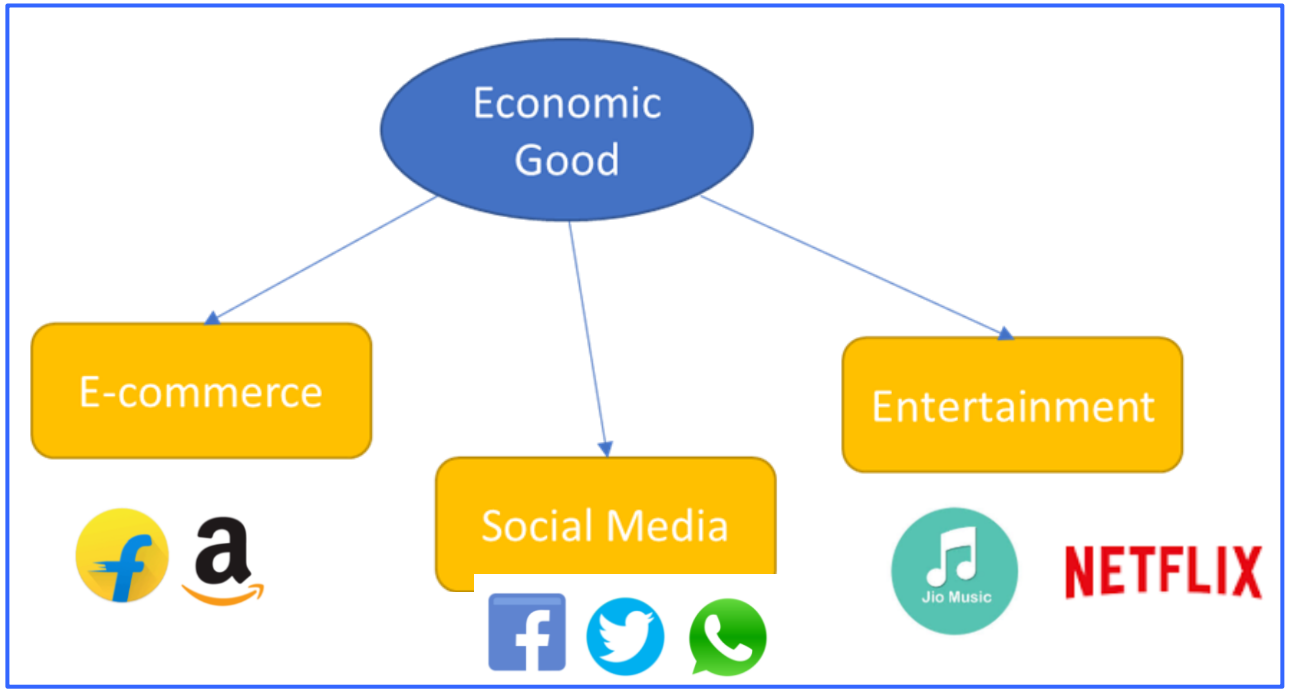


Internet User Base in India (in million)

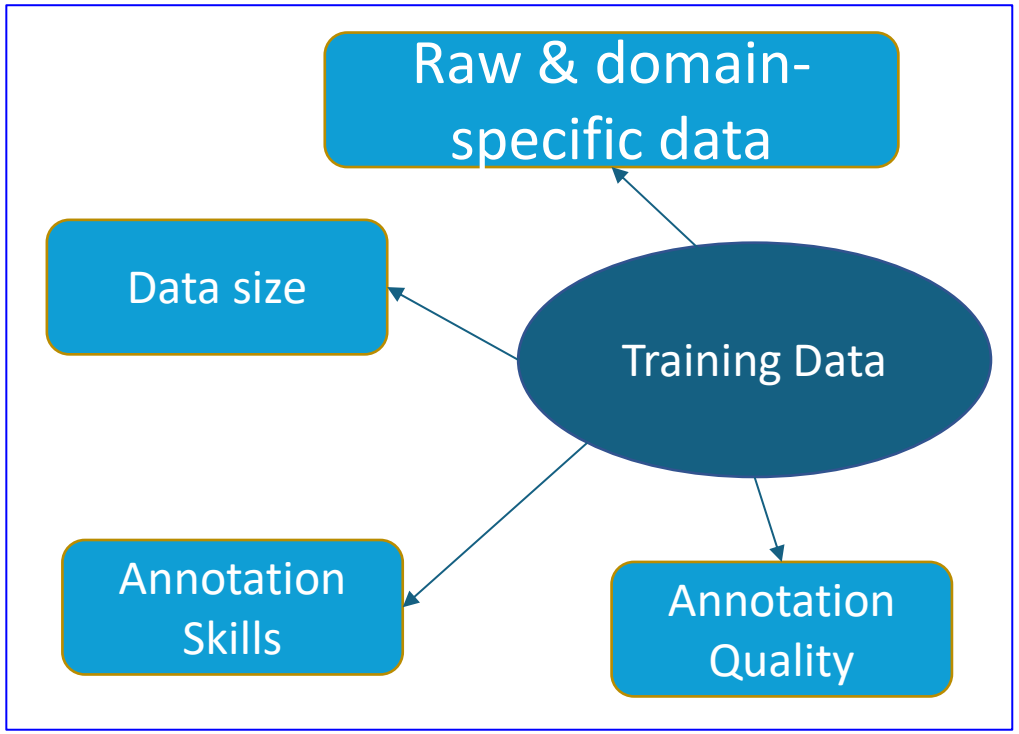
Source: Indian Languages: Defining India's Internet KPMG-Google Report 2017


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

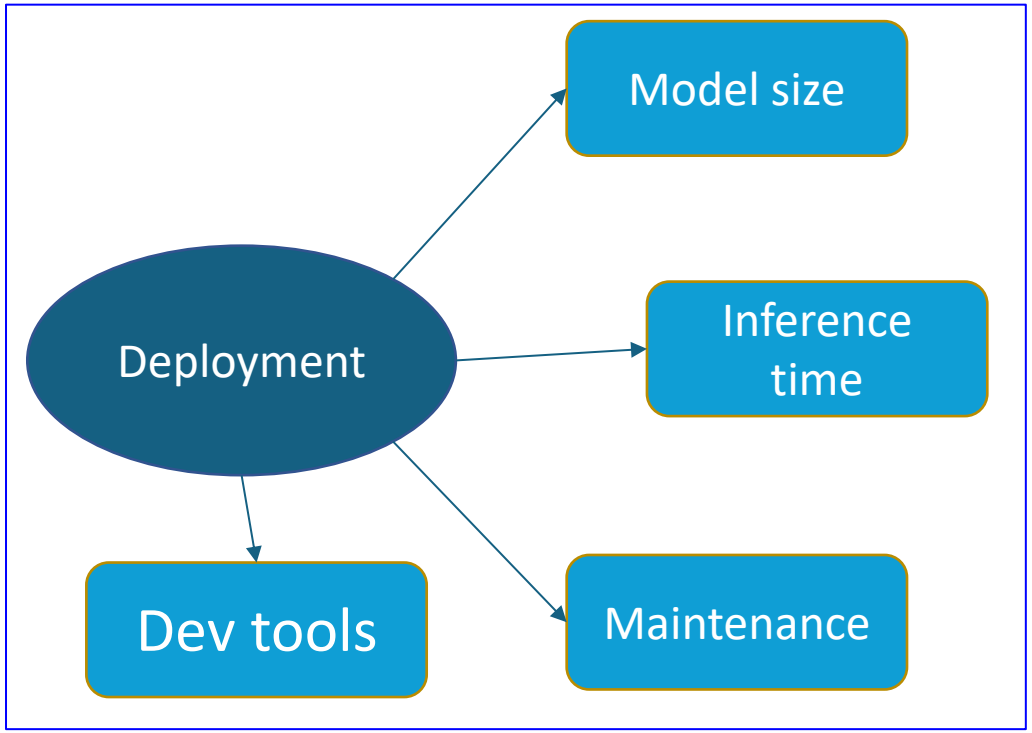

ENGLISH
ITALIANO
DEUTSCH
PORTUGUES
FRANCAIS
ESPANOL
People-People Contact
Cultural Exchanges
Easier Travel and Migration


Economic Good
E-commerce
Social Media
Entertainment
Jio Music
NETFLIX

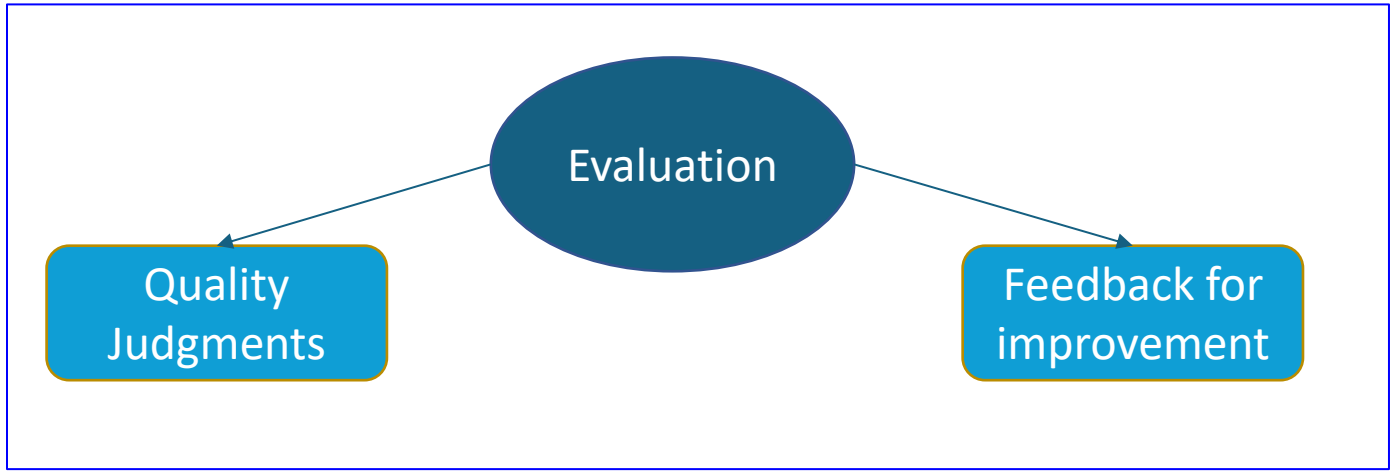
Scalability Challenges for NLP solutions



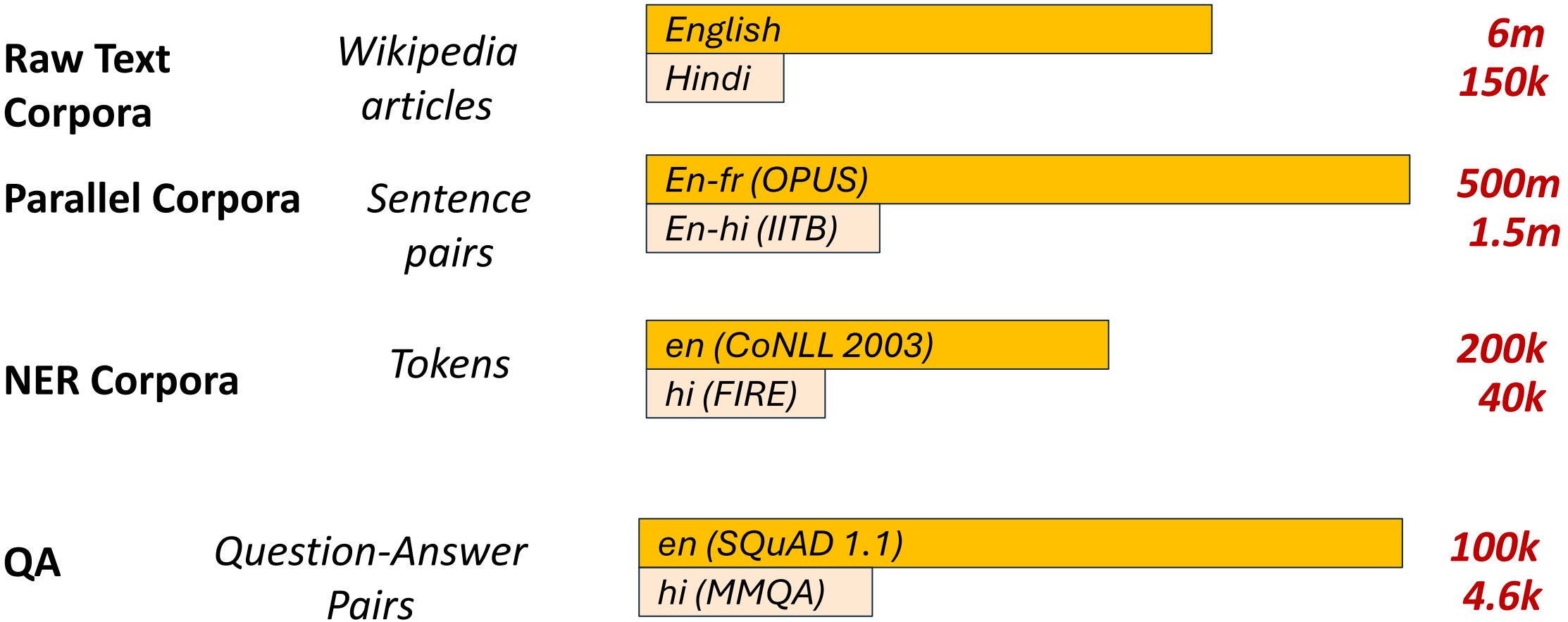
Languages



Effort and cost increase as languages increase



We are faced with a huge data skew

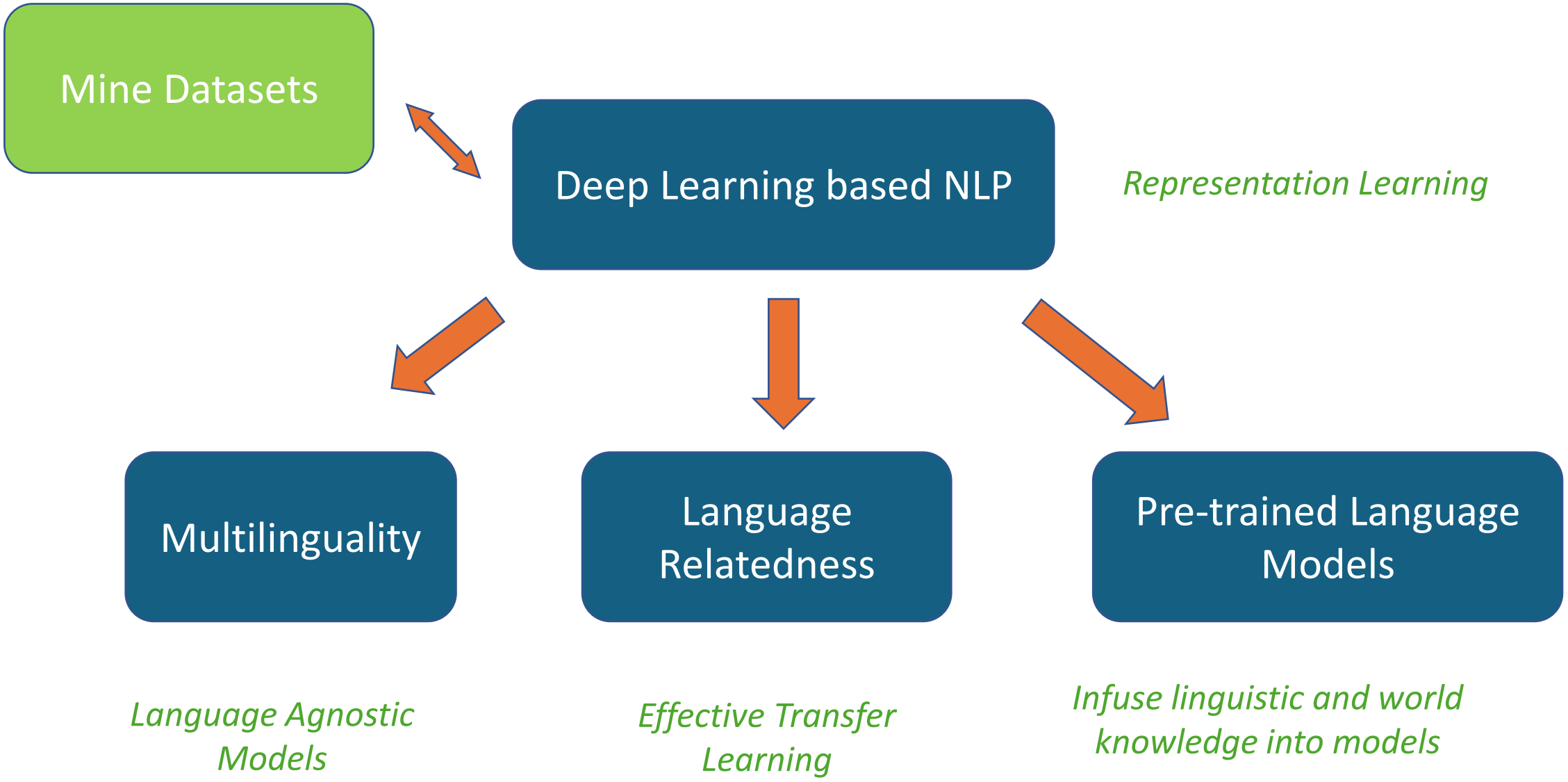


How do we approach this scalability problem?

Each generation of NLP technology takes a big step in addressing the scalability challenge

Rule-based		• Language-specific knowledge encoding • Brittleness • Linguistic Expertise
Classical ML	• More robust • Lesser Linguistic expertise	• Lot of labelled data required • Feature Engineering resources and effort
Deep Learning	• Automatic Feature Learning • Better semantic representation • Multilingual modeling • Better Generation capabilities	• Still need a good amount of labelled data
Pre-training	• Reduce labelled data requirements	• … but you still need labelled data • Task-specific models
LLMs	• In-context learning • Data generators • Unseen tasks/task-composition • Open-ended tasks	• Evaluation of non-trivial tasks • Alignment Data

How do we approach this scalability problem?



Data Curation

Build on top of community efforts!

	en-as	en-bn	en-gu	en-hi	en-kn	en-ml	en-mr	en-or	en-pa	en-ta	en-te	Total
JW300	46	269	305	510	316	371	289	-	374	718	203	3400
banglanmt	-	2380	-	-	-	-	-	-	-	-	-	2380
iitb	-	-	-	1603	-	-	-	-	-	-	-	1603
cvit-pib	-	92	58	267	-	43	114	94	101	116	45	930
wikimatrix[5]	-	281	-	231	-	72	124	-	-	95	92	895
OpenSubtitles	-	372	-	81	-	357	-	-	-	28	23	862
Tanzil	-	185	-	185	-	185	-	-	-	92	-	647
KDE4	6	35	31	85	13	39	12	8	78	79	14	402
PMIndia V1	7	23	42	50	29	27	29	32	28	33	33	333
GNOME	29	40	38	30	24	23	26	21	33	31	37	332
bible-uedin	-	-	16	62	61	61	60	-	-	-	62	321
Ubuntu	21	28	27	25	22	22	26	20	29	25	24	269
ufal	-	-	-	-	-	-	-	-	-	167	-	167
sipc	-	21	-	38	-	30	-	-	-	35	43	166
GlobalVoices	-	138	-	2	-	-	-	326	1	-	-	142
TED2020	< 1	10	16	46	2	6	22	-	752	11	5	120
Mozilla-I10n	7	21	-	< 1	12	13	15	8	-	17	25	119
odiencorp 2.0	-	-	-	-	-	-	-	91	-	-	-	91
Tatoeba	< 1	5	< 1	11	< 1	< 1	53	< 1	< 1	< 1	< 1	71
urst	-	-	65	-	-	-	-	-	-	-	-	65
alt	-	20	-	20	-	-	-	-	-	-	-	40
mtenglish2odia	-	-	-	-	-	-	-	35	-	-	-	35
nlpc	-	-	-	-	-	-	-	-	-	31	-	31
wmt-2019-wiki	-	-	18	-	-	-	-	-	-	-	-	18
wmt2019-govin	-	-	11	-	-	-	-	-	-	-	-	11
tico19	-	< 1	< 1	< 1	< 1	< 1	< 1	-	< 1	< 1	< 1	6
ELRC_2922	-	< 1	-	< 1	-	< 1	-	-	-	< 1	< 1	1
Total	108	3496	611	2818	472	1237	758	229	631	1456	593	**12408**

Just compiling existing corpora helped build models outperforming existing publicly available models!

Cataloging is a useful exercise

OPUS
OPEN PARALLEL CORPORA

The Indic NLP Catalog

Glot500 Corpus

Indonesian NLP Data Catalogue

Monolingual Data Collection

Compile the collective knowledge of the web!

IndicCorp v1

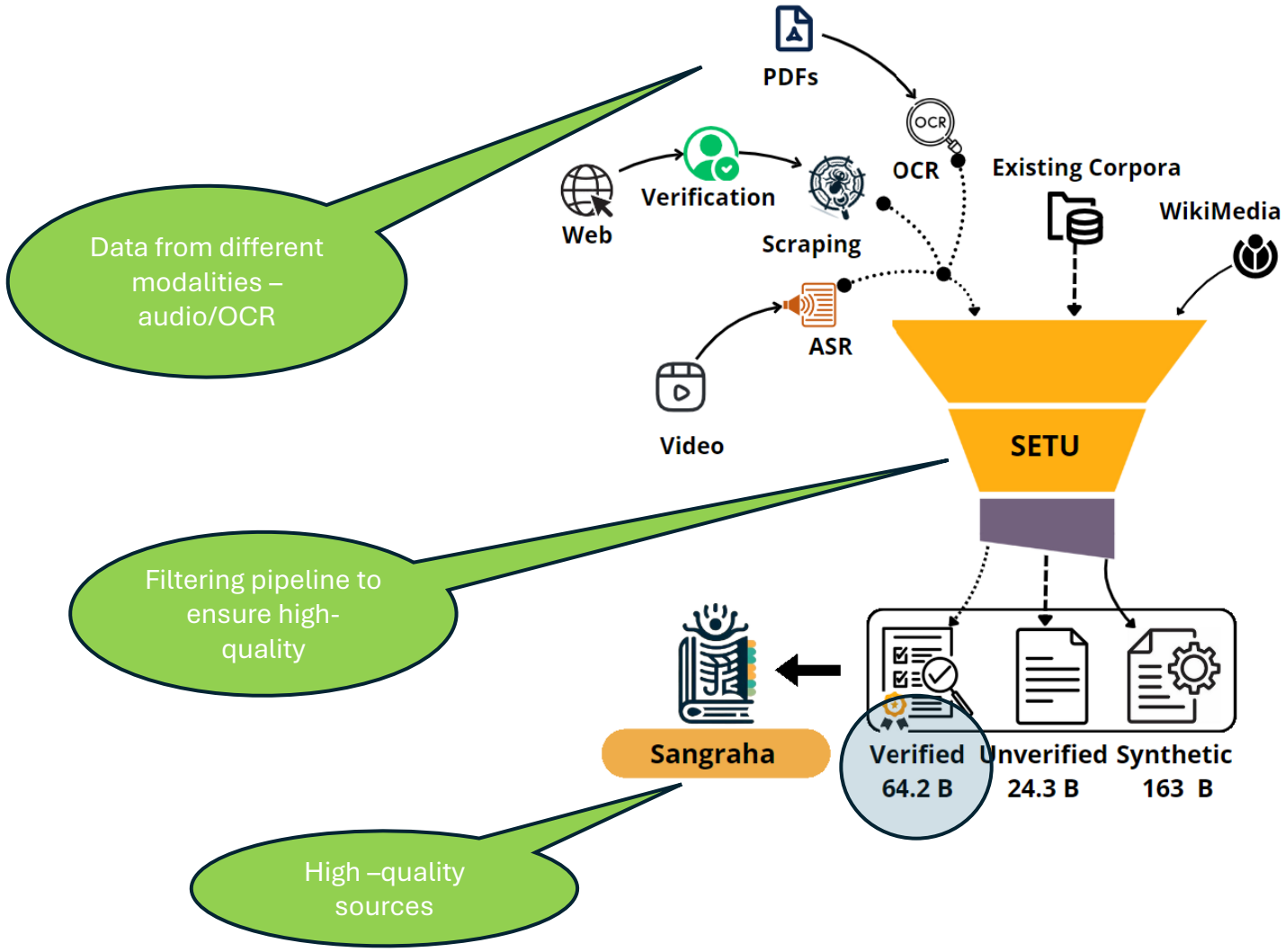
Sentence-level
Web-sources

IndicCorp v2

Larger corpora
Larger language coverage

Sangraha

Document level
Diverse sources
Better filtering



THE Key Resource

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

LM Training Corpora

Labelled Data Mining

Harness the wisdom of the crowds!

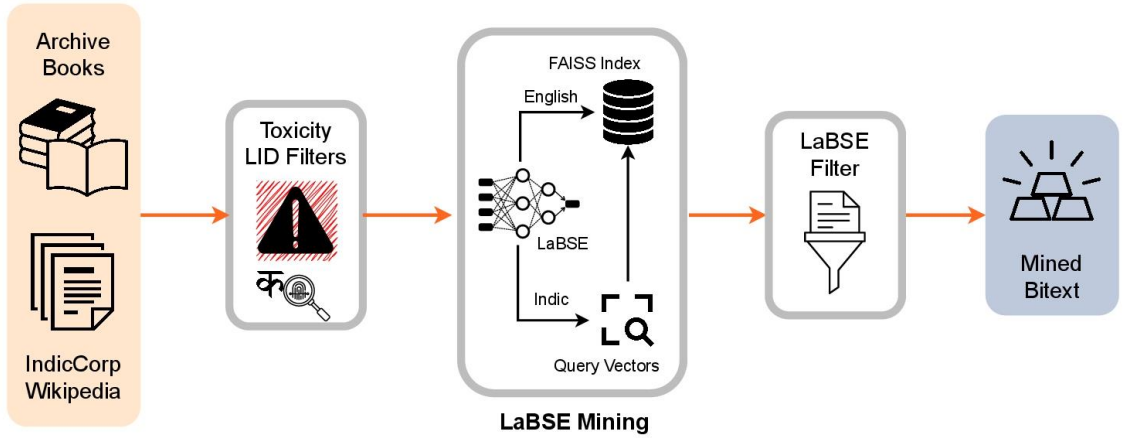
Mine Parallel Corpora

Heuristic-based

Content-based

Sentence-level

Document-level



200 million sentence pairs

1.3 million document pairs

Mined data far exceeds dedicated annotated data

Most significant contributor to translation quality improvement

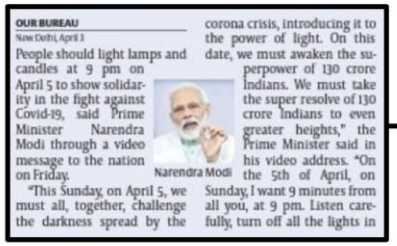
Filtering necessary to ensure high-quality and safe content

Creativity is the limit for mining data of different kinds!

BIOGRAPHY GENERATION

→ Murmu assumed office on 25th July and succeeded Ram Nath Kovind

HEADLINE GENERATION



OUR BUREAU
New Delhi, April 3

People should light lamps and candles at 9 pm on April 5 to show solidarity in the fight against Covid-19, said Prime Minister Narendra Modi through a video message to the nation on Friday.

"This Sunday, on April 5, we must all, together, challenge the darkness spread by the corona crisis, introducing it to the power of light. On this date, we must awaken the superpower of 130 crore Indians. We must take the super resolve of 130 crore Indians to even greater heights," the Prime Minister said in his video address. "On the 5th of April, on Sunday, I want 9 minutes from all you, at 9 pm. Listen carefully, turn off all the lights in

Narendra Modi

→ Dispel the virus darkness with light: PM to people

SENTENCE SUMMARIZATION

India 's financial markets are closed on Monday for a public holiday.

→ India markets closed for holiday

PARAPHRASE GENERATION

दिल्ली विश्वविद्यालय, भारत में उच्च शिक्षा के लिए एक प्रतिष्ठित संस्थान है।

→ A आ

The University of Delhi is a prestigious institution for higher education in India.

Delhi University is one of the famous universities of the country.

QUESTION GENERATION

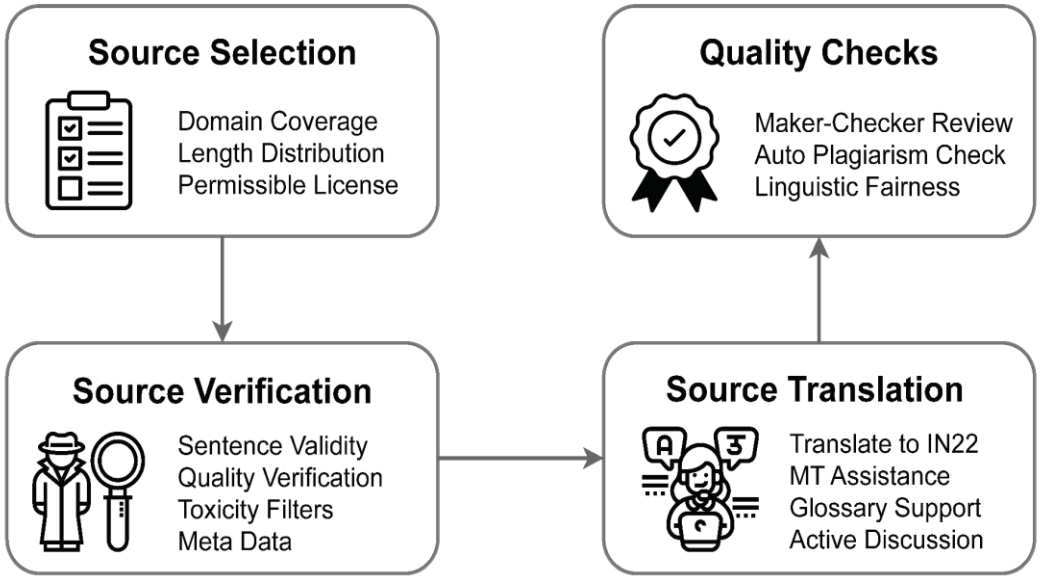
SQuAD → A आ → HI SQuAD

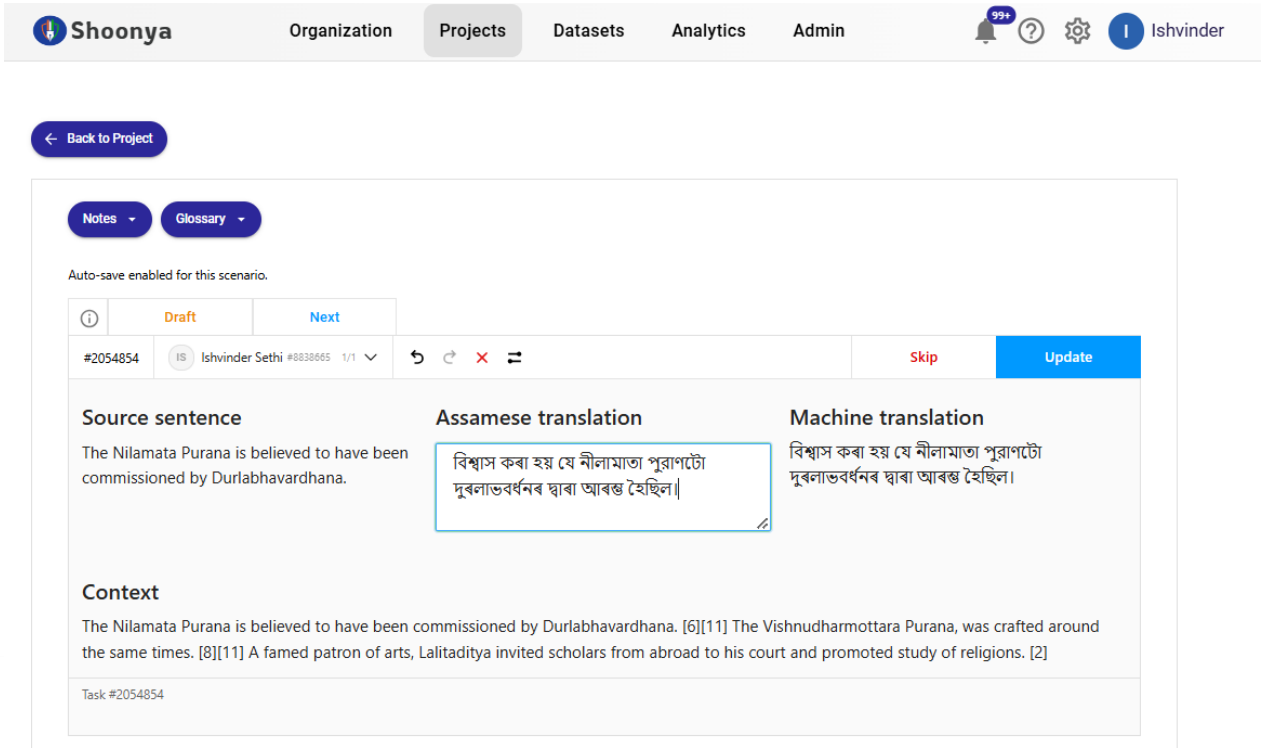
INDICTRANS

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

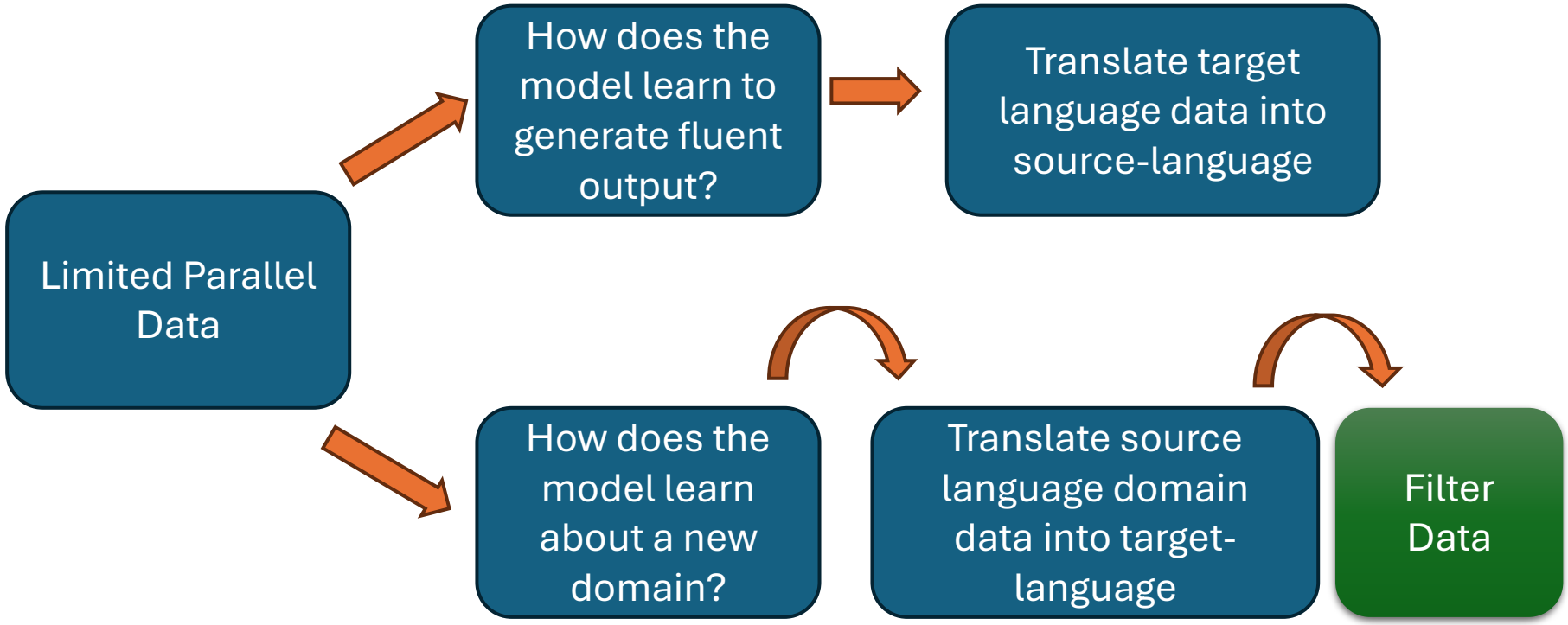




- Need processes in place to ensure high quality
- Provide tools to make translators productive

Synthetic Data Generation

Model generated data to address specific phenomena



Significantly boosts quality for Low-resource languages

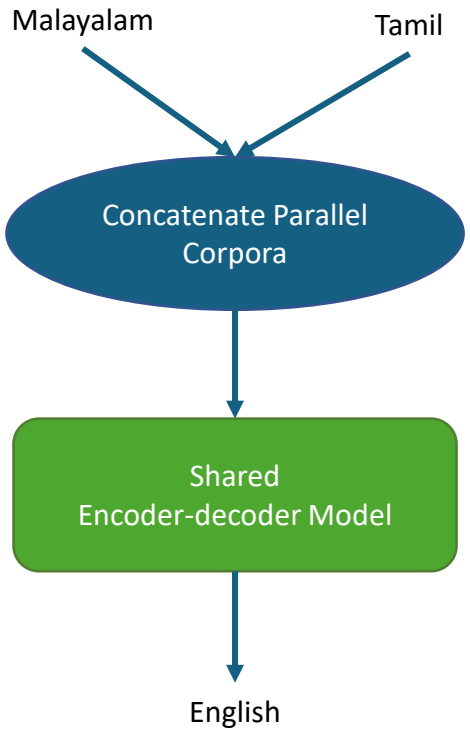
Can make models domain-aware, handle natural source language text

***Risk**: Regressions possible*

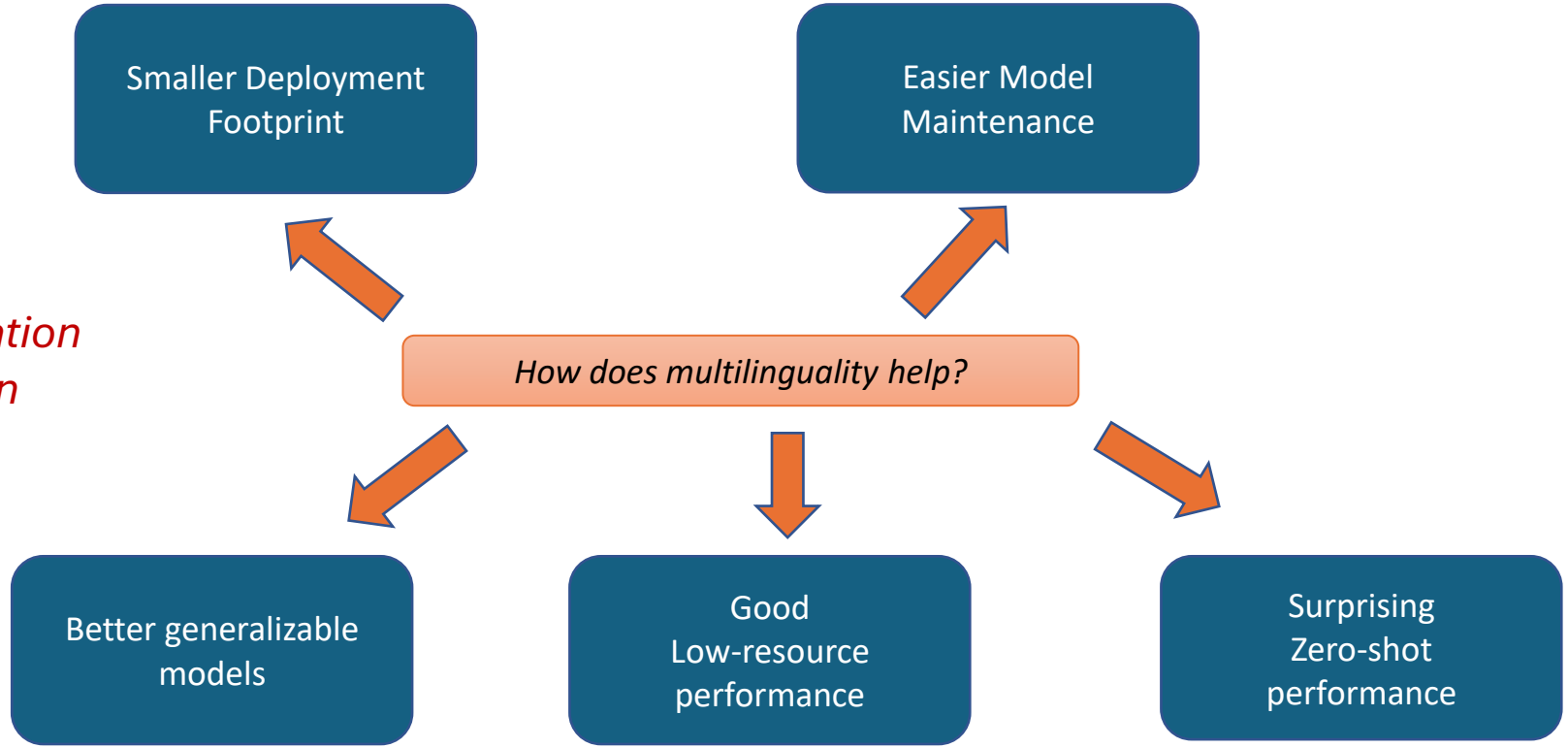
Multilingual Modeling

Multilingual models are now a no-brainer!

Let the rich languages help their poor cousins!



Script unification
Romanization



Significant improvement for low-resource languages

In the era of English-dominant LLMs, how do we best transfer from knowledge from English?

Language Model Pre-training

IndicBERT ***IndicBART***

Infuse linguistic and world-knowledge into models

Challenges with Massive Pretrained models

- Limited Data Coverage
- Tokenizer Representation
- Limited Language Coverage

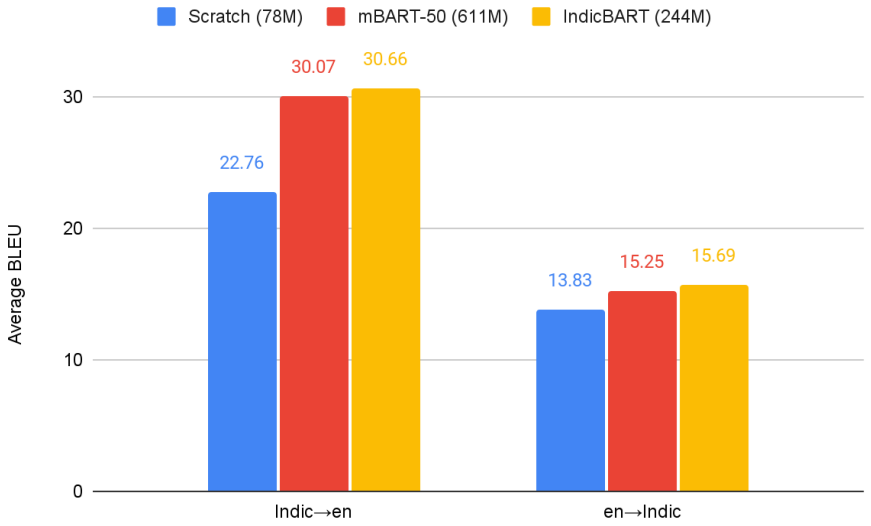
Challenges with Building India-specific models

- Compute!
- Pre-training of massive English corpora is valuable

Language model adaptation

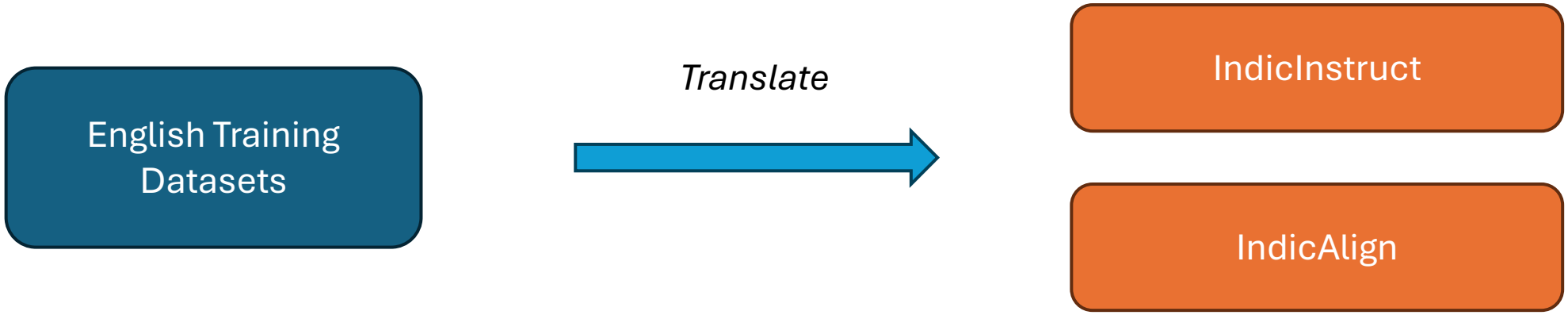
Vocabulary adaptation

Avoiding catastrophic forgetting



Models	Classification					Structure Prediction		QA	Retreival
	Indic Sentiment	Indic XNLI	Indic COPA	Indic XPara.	MASSIVE (Intent)	Naama-Padam	MASSIVE (Slotfill)	Indic QA	FLORES
IndicBERT v1	61.8	42.8	51.0	47.5	-	25.3	-	10.1	1.1
mBERT	69.5	54.7	51.7	55.2	13.2	63.0	6.2	32.9	32.3
XLMR	84.0	69.7	60.1	56.7	66.6	71.7	50.0	44.8	3.1
MuRIL	85.1	72.4	58.9	**60.8**	77.2	**74.3**	57.0	48.3	52.3
v1-data	85.7	66.4	52.4	49.6	25.8	58.3	34.4	37.6	54.9
IndicBERT v2	**88.3**	73.0	62.7	56.9	78.8	73.2	56.7	47.7	69.4
+Samanantar	**88.3**	74.3	**63.0**	57.0	78.8	72.4	**57.3**	49.2	64.7
+Back-Trans.	87.5	69.7	53.8	50.7	77.4	71.9	54.6	42.2	68.6
IndicBERT-SS	88.1	**73.9**	64.2	56.4	**80.7**	66.6	**57.3**	**49.7**	**71.2**

Machine Translation as an enabler to scaling



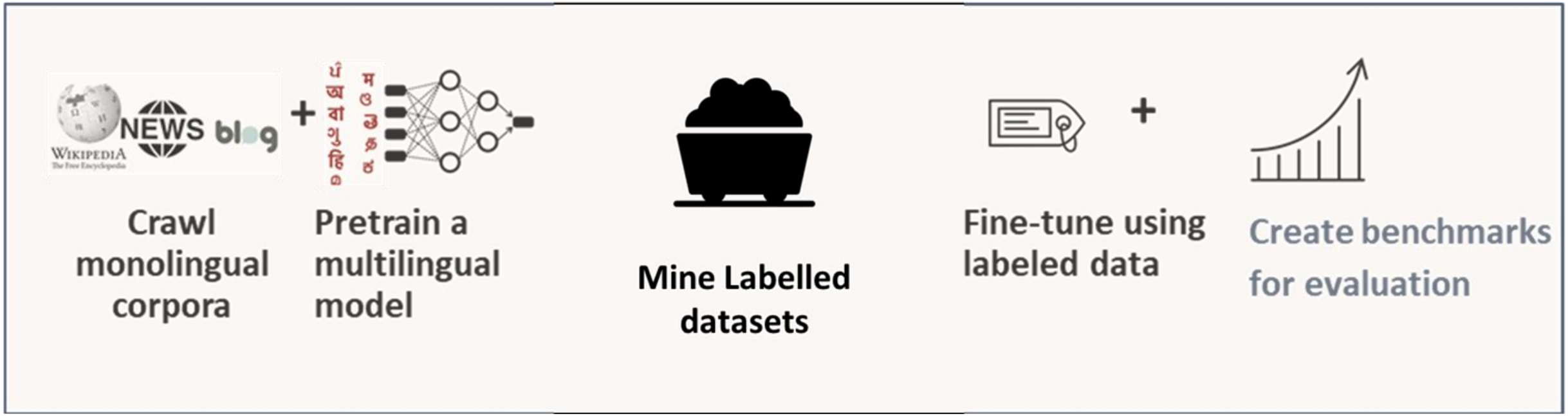
- *Wide variety of datasets available in English*
- *MT generated training data more relevant for decoder-only models*

Are MT generated benchmarks good?

- Quality issues may result in misleading results
- Applicable only when translation quality is reasonably good

- Human generated benchmarks for low-resource languages might actually be worse!

The “Recipe” for Language Scalability



Interesting Directions

- Better Cross-lingual transfer from English for decoder LLMs
- Synthetic SFT/Preference Data Generation
- Scaling Evaluation: Learned Metrics, LLM based Evaluation

Thank you!

anoop.kunchukuttan@gmail.com

https://anoopkunchukuttan.gitlab.io/