



Advances in Multilingual Machine Translation and Evaluation for Indian Languages

Raj Dabre
ASTREC, NICT Japan

prajdabre@gmail.com

02/09/2024

Link to talk!

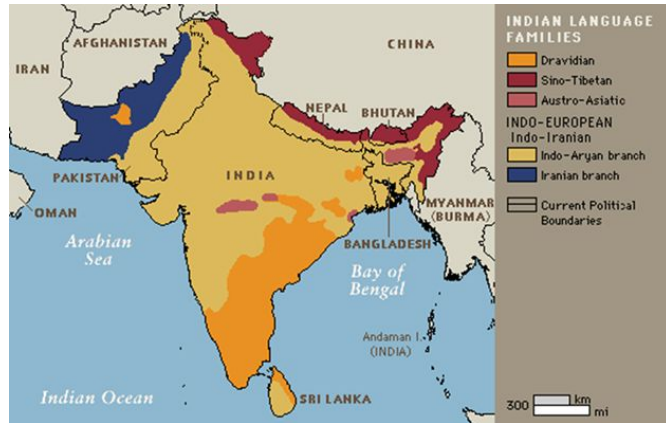


About Me: Raj Dabre (raj.dabre@nict.go.jp)

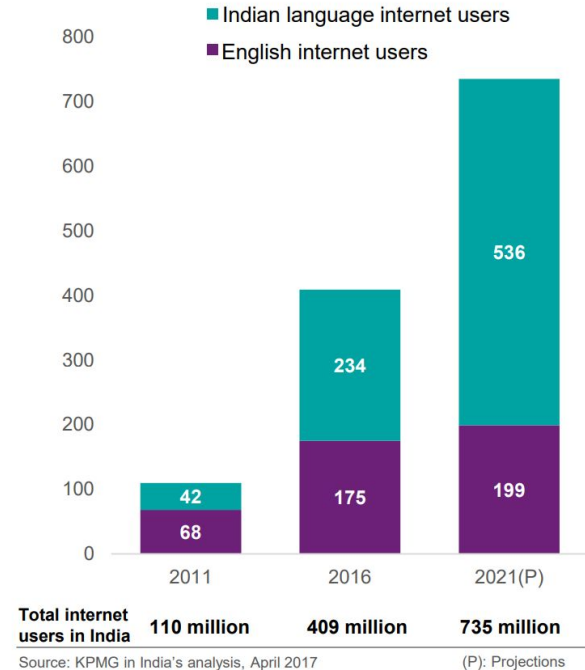
- Experience
 - 2018-present: Researcher at NICT, Japan
 - Adjunct Faculty at IIT Madras
 - Visiting Assistant Professor at IIT Bombay
 - 2014-2018: MEXT Ph.D. scholar at Kyoto University, Japan (SMT to NMT)
 - 2011-2014: M.Tech. Government RA at IIT Bombay, India (RBMT to SMT)
- Research
 - Low-Resource Natural Language Processing
 - **Multilingual Machine and Speech Translation: 2012-present**
 - **Document Level Machine Translation: 2021-**
 - **Large Scale Pre-training for Generation: 2021-**
 - Efficient Deep Learning:
 - **Compact, flexible and fast models (2018-present)**

Indian Languages MT: *But Why?*

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
- Leading economy



Internet User Base in India (in millions)

Indic MT Is *Crucial* But Is **Data Hungry**

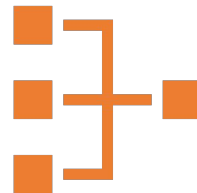
What Is (was) Missing?



Pretraining
Data and
models



MT Training
Data

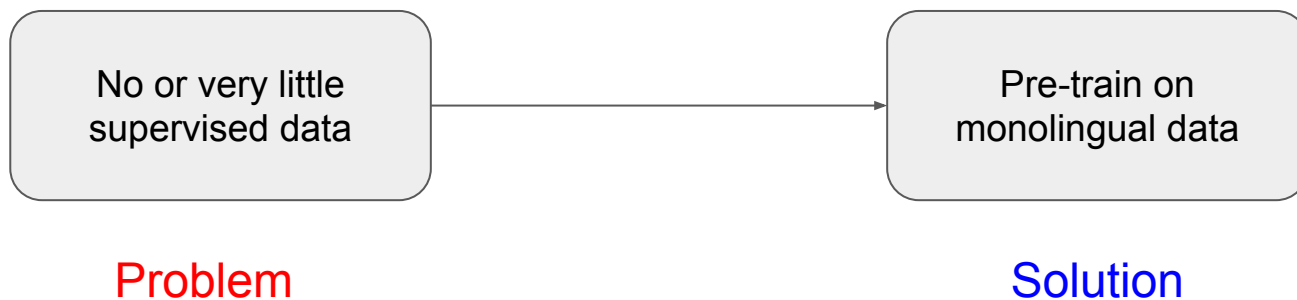


MT Models and
Evaluation

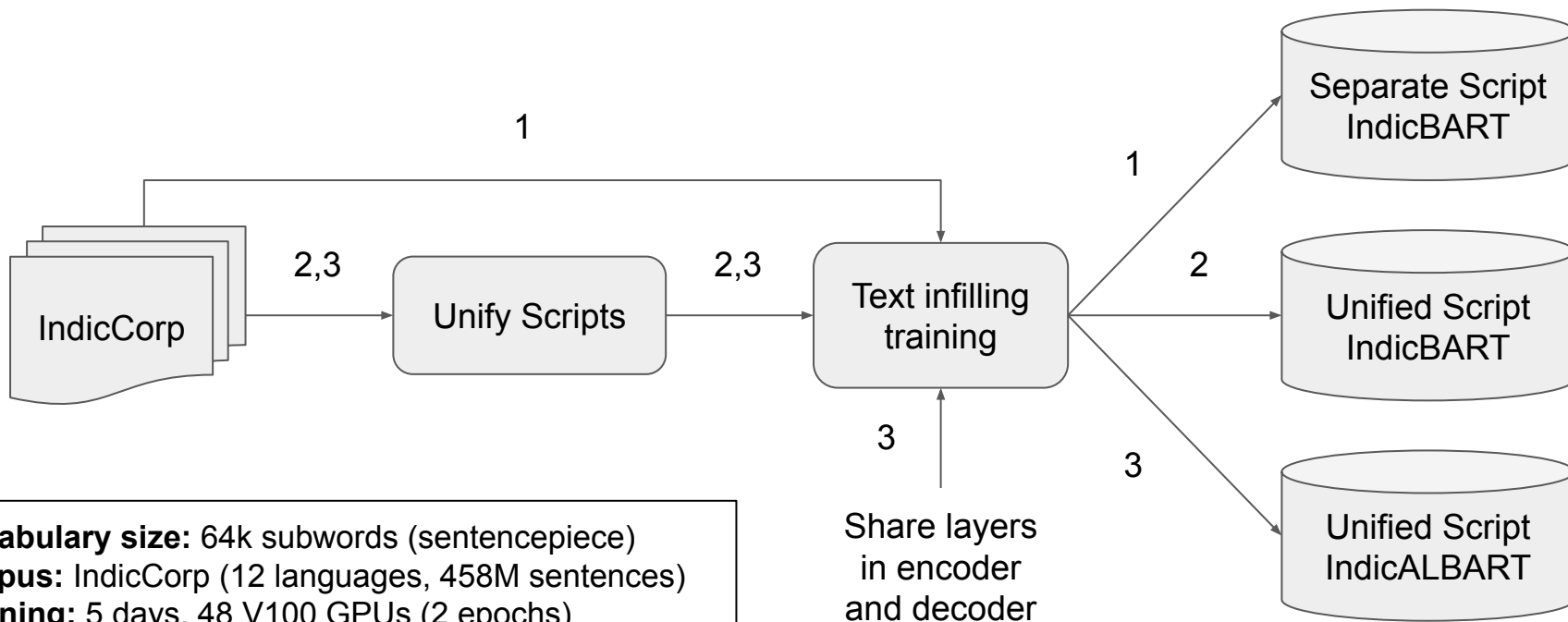
This Talk

- IndicBART
 - First ever Indic specific NLG pre-trained models and datasets
- IndicTrans2
 - Current SOTA MT model for 22 Indic languages
- IndicLLM Suite
 - Monolingual data mining and synthesis
- MT using LLMs and Prompting
 - Use feature scoring, monotonic word ordering, transliteration
- Indic MT Evaluation
 - Reliable evaluation of Indic MT and its limitations

IndicBART ([Dabre+, 2021](#))



IndicBART Models

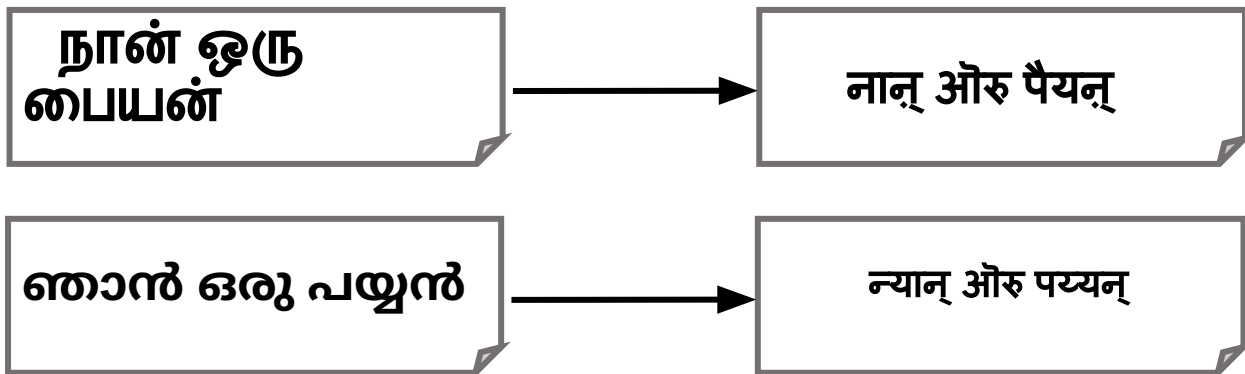


Vocabulary size: 64k subwords (sentencepiece)
Corpus: IndicCorp (12 languages, 458M sentences)
Training: 5 days, 48 V100 GPUs (2 epochs)
Params: 97M (IndicALBART), 244M (IndicBART)

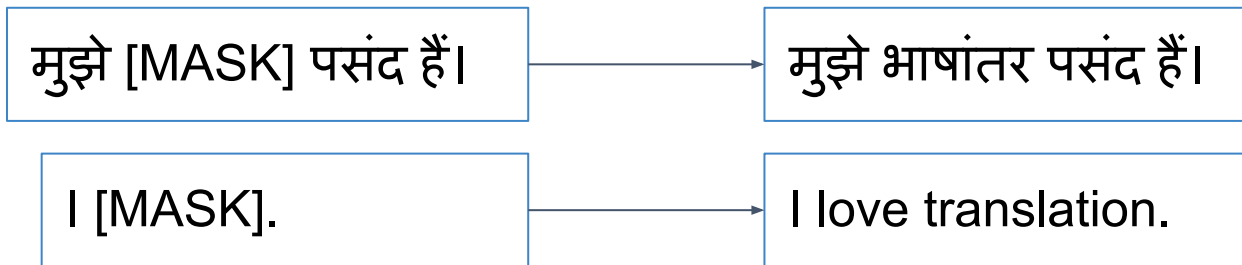
mBART: 250k subwords, 25-50 languages, over 2 weeks on 256 GPUs, 611M params

Script Unification

- Many languages need large vocabulary
- Script unification by converting to Devanagari
 - Increased vocabulary sharing
 - Compact vocabularies for compact models



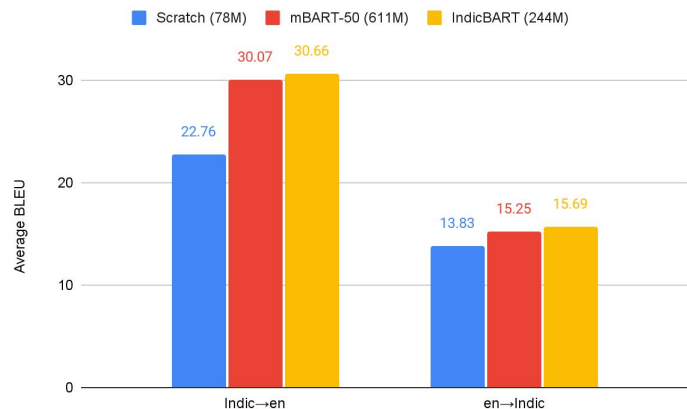
What is infilling?



Downstream Task: Machine Translation

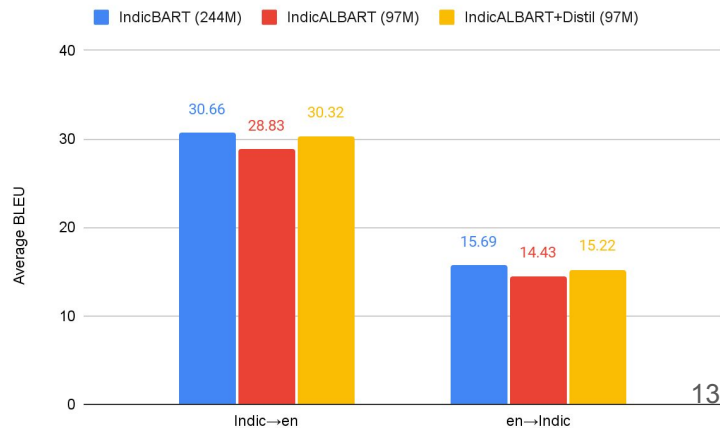
Pre-training Results

- **Large impact of pre-training**
 - *Indic*→*En*: 22.76→30.66
 - *En*→*Indic*: 13.83→15.69
- **Indic**→**En** gains more than **En**→**Indic**



Parameter Tying (*IndicALBART*)

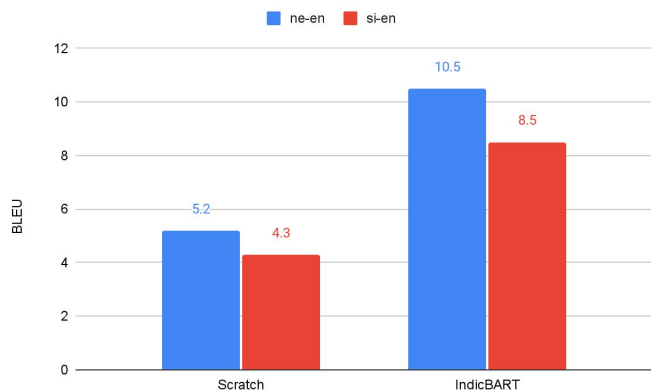
- Significant compression and modest drop in BLEU
- **Drop can be mostly recovered with distillation**
 - *IndicALBART* improves 0.8 to 1.5 BLEU
 - *Distilled IndicALBART* ≈ *IndicBART*
- **Scratch model performance also improves**
 - *Indic*→*En*: 22.76→29.11
 - *En*→*Indic*: 13.83→15.33



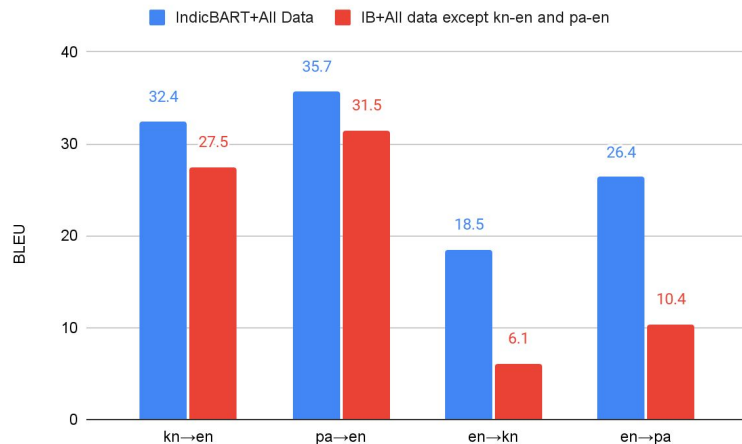
Results on WAT 2021 MultiIndicMT dataset
(10 language pairs, 326K sentence pairs, N-way development and test sets)

Downstream Task: Machine Translation (Zero Shot)

- Unseen languages during pre-training
 - Nepalese and Sinhala
 - Map scripts to Devanagari
 - Same data as in Liu et al. (2020)
 - FLORES evaluation sets
- **Result:**
 - *IndicBART effective on unseen languages*
 - *4-5 BLEU gains*
 - Liu et al. (2020) has seen both languages

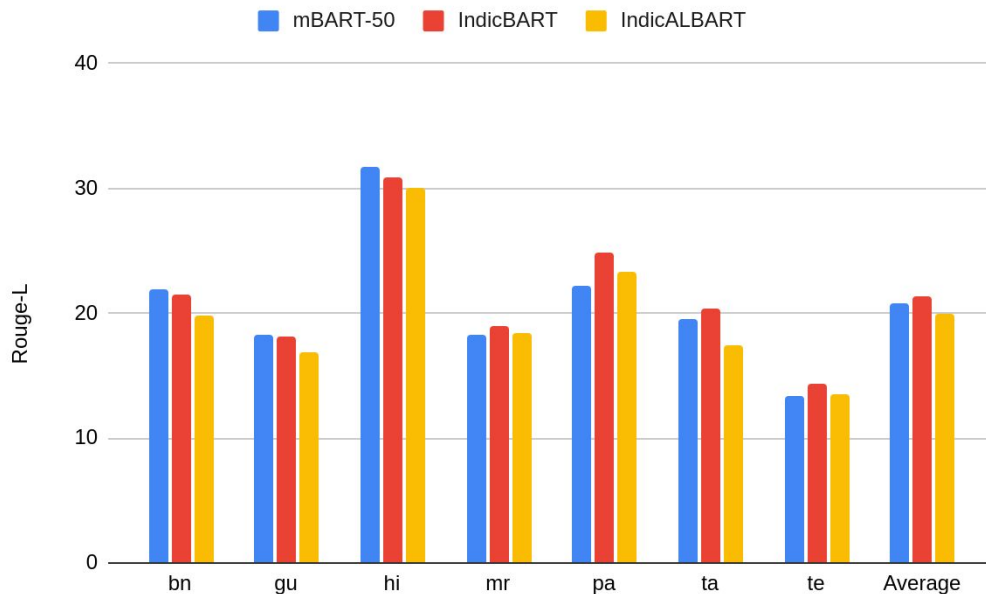


- Unseen languages during fine-tuning
 - WAT 2021 dataset (PMI for training)
 - Fine-tune on all except Kn-En, Pa-En
- **Result:**
 - *Indic→En: 4-5 BLEU below all data FT*
 - *Due to language relatedness?*
 - *En→Indic is the real challenge*



Not MT But: Side note on Summarization

- 7 Indic languages subset from XL-Sum
 - Hasan et al. (2021)
- Results:
 - **IndicBART > mBART-50 on average**
 - *IndicALBART not far behind*
 - **mBART-50 better for Hi, Bn, Gu**
 - *Impact of pre-training data?*
 - **IndicBART better for rest**
 - *Pa unseen by mBART-50*
- Impact of distillation?
 - *Can IndicALBART improve?*
- Also see:
 - IndicNLG Benchmark ([Kumar+, 2022](#))



But ***SOTA Models*** Need **Much More!**

IndicTrans2 ([Gala+, 2023](#))

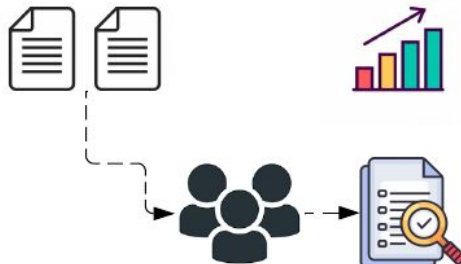
Contributions



High Quality Data

232 M
Mined
sentences

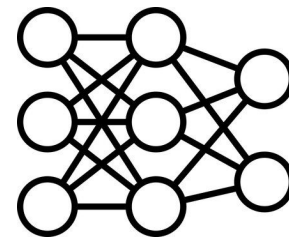
800 K
Seed
sentences



Robust, manually
created Benchmarks.

IN22-Gen
1st India-centric
multi-domain
benchmark

IN22-Conv
1st
Conversation
Translation
benchmark



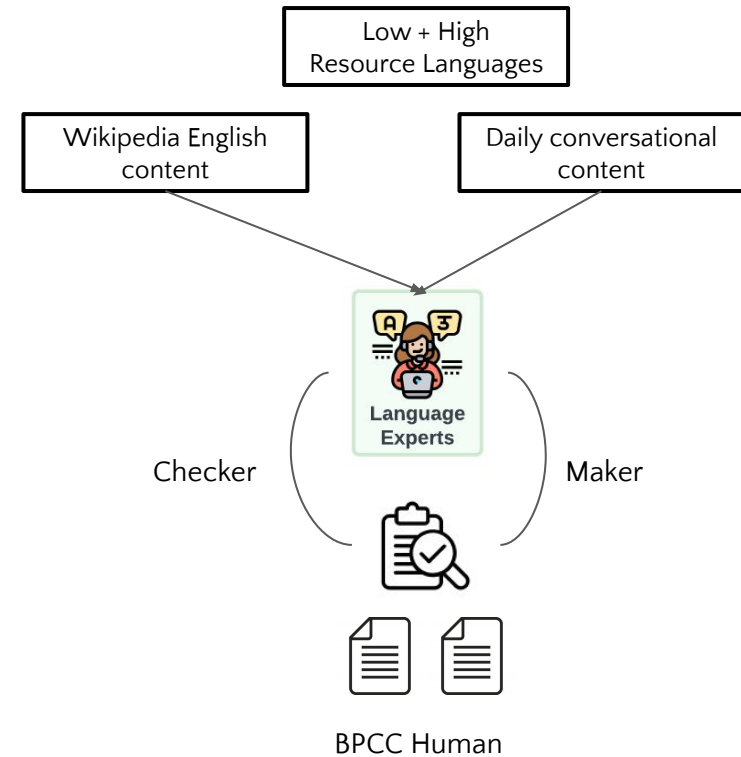
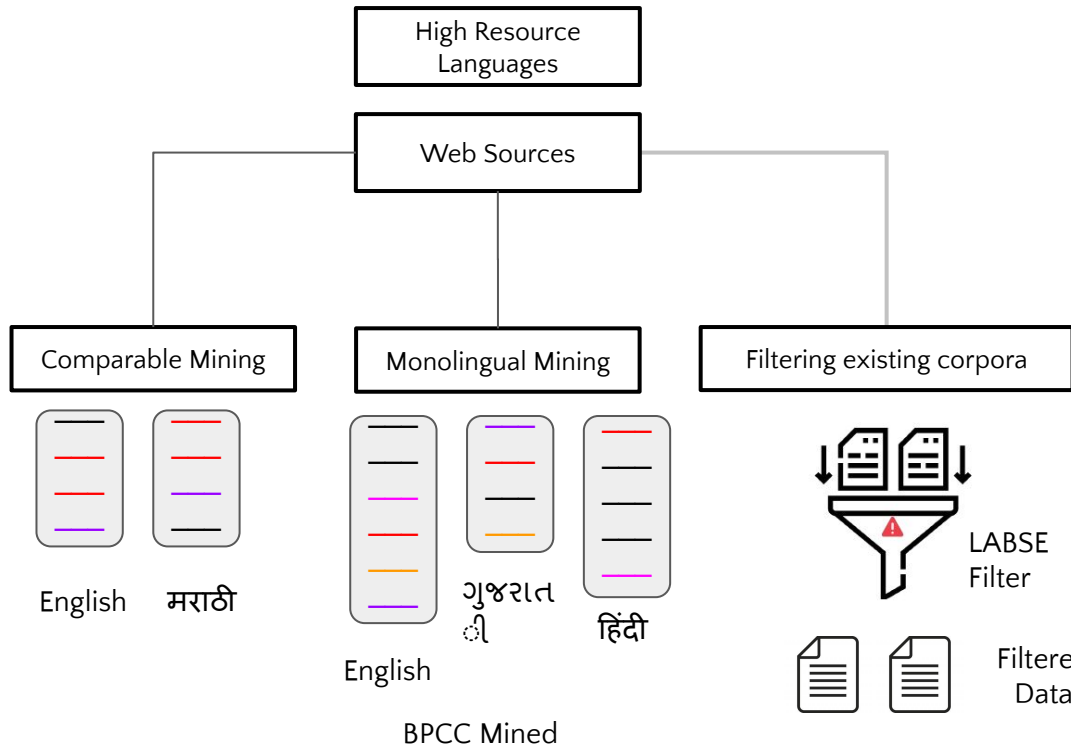
IndicTrans2

SOTA
SOTA
open-source
models for Indic
languages.

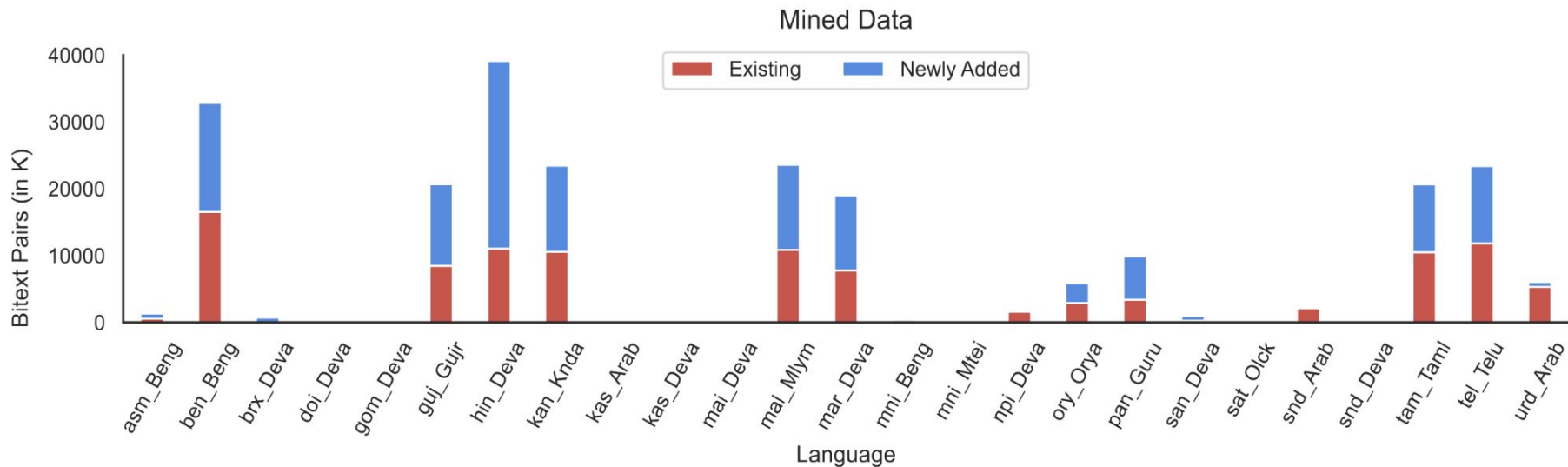
#1
First model that
supports all 22
scheduled
Indian
languages.

MIT License

How did we address the data problem ?

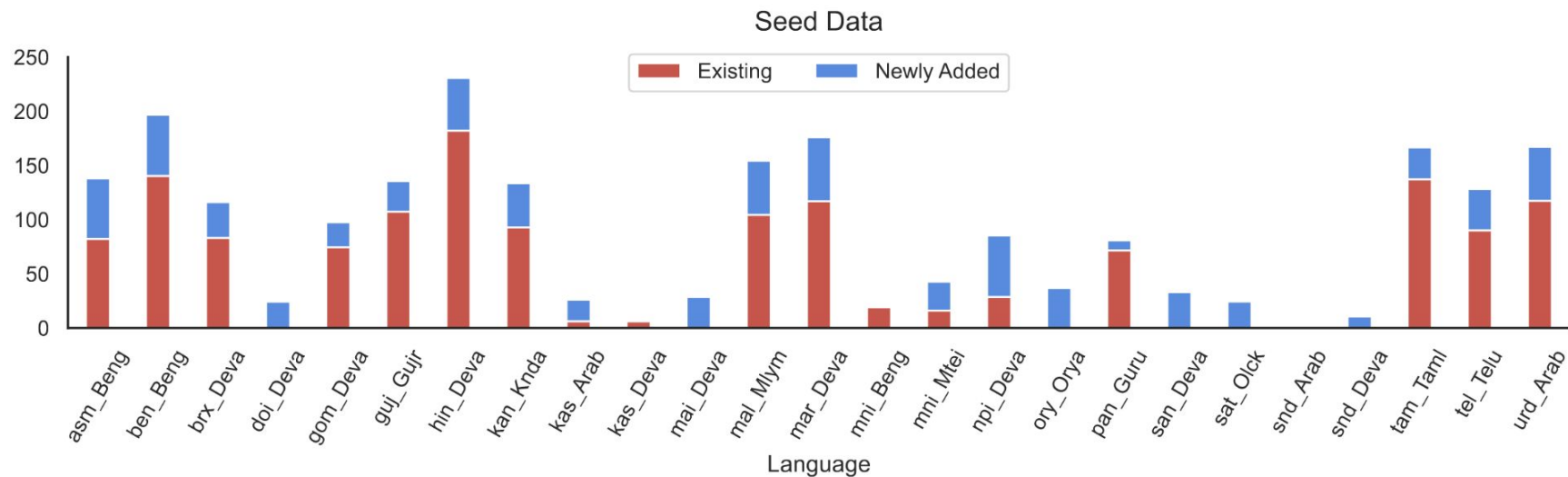


Our Data Contributions : Mined Data



of bitext pairs = -126M (high filter threshold)

Our Data Contributions : Seed Data



of bitext pairs = ~800K

IndicMT Benchmark: Why yet another MT Benchmark?

- ✗ No India-centric content
- ✗ No test set for all 22 languages
- ✗ Modern use-cases (Conversations)
- ✗ Limited domain & length diversity
- ✗ Limited source diversity

Existing Benchmarks:

- FLORES-200
- NTREX-128
- WMT, WAT



IN22
n-way
multi-domain
supports all 22
scheduled Indian
languages

IN22-Gen

1024 sentences
13 domains

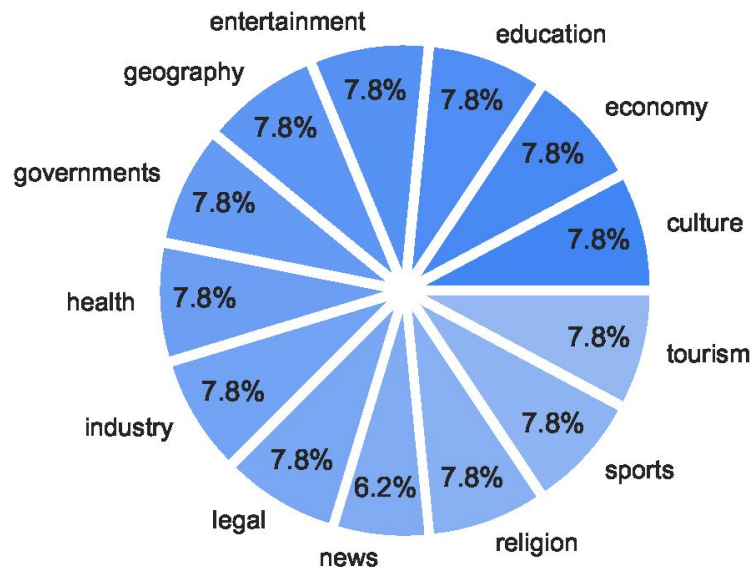
IN22-Conv

1503 sentences
44

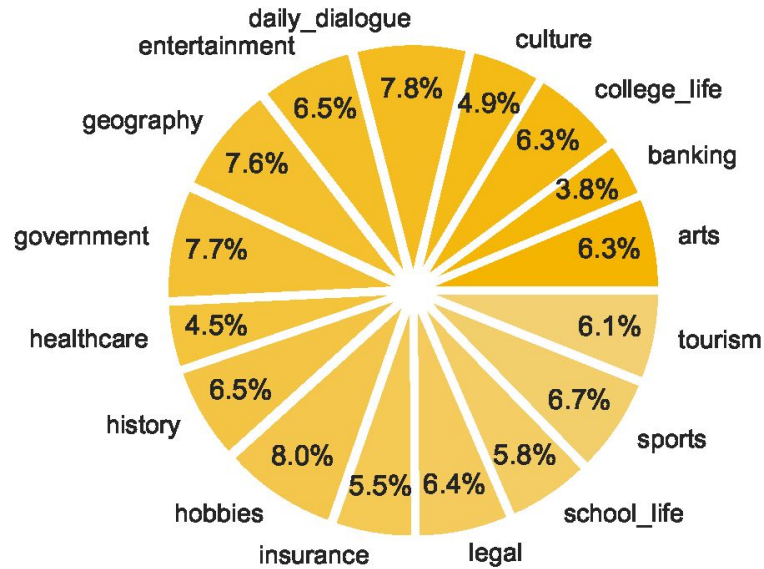
conversations
Useful for evaluating
document-level
translation capabilities
of LLMs

IN22 Benchmark : Domain Diversity

IN22-Gen

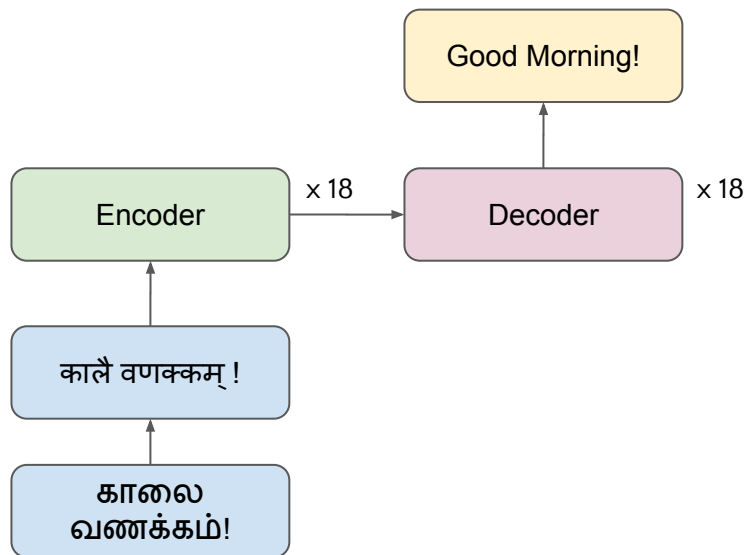


IN22-Conv

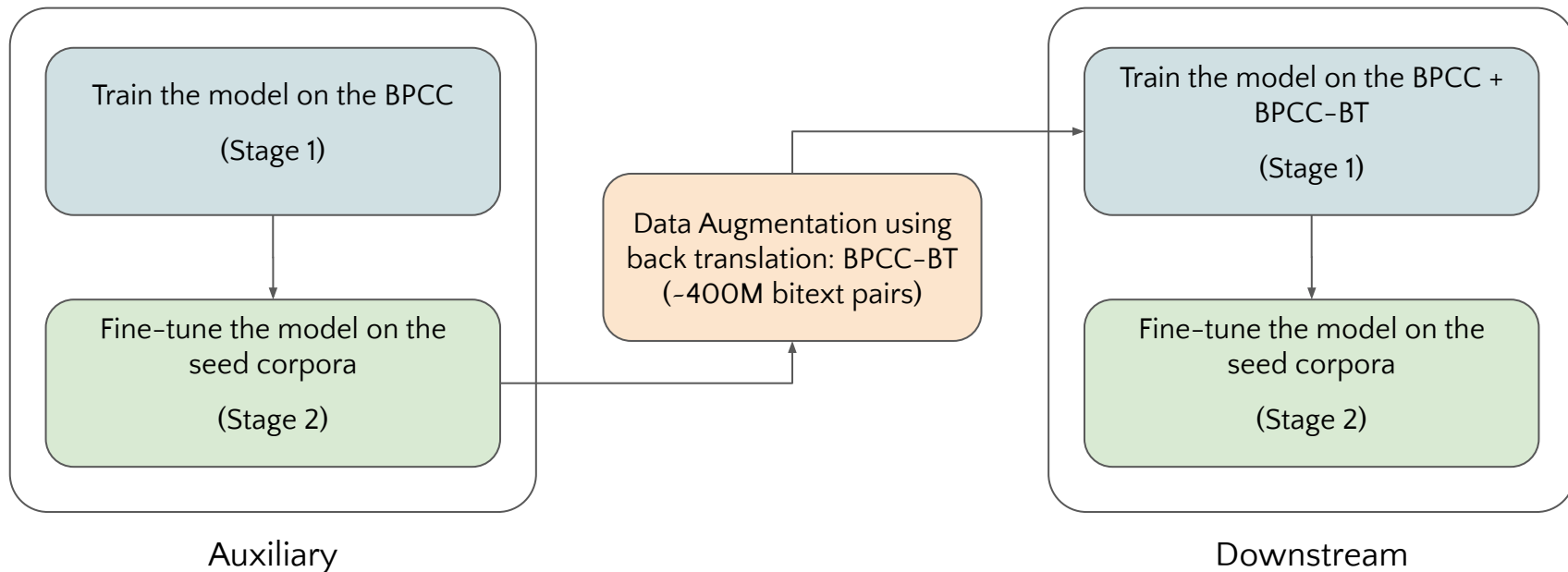


Training IndicTrans2

- Joint multilingual model for 22 languages.
- Supports 25 language script combinations.
- Script-sharing wherever feasible to enable transfer learning.
- 2 Models : En-XX and XX-En
- -1.1 Billion parameter models.
- 18 Encoder Layers, 18 Decoder Layers, 16 attention heads / layer.
- Deeper Models >> Wider Models
- FFN_dim 8192
- Embedding_dim 1024

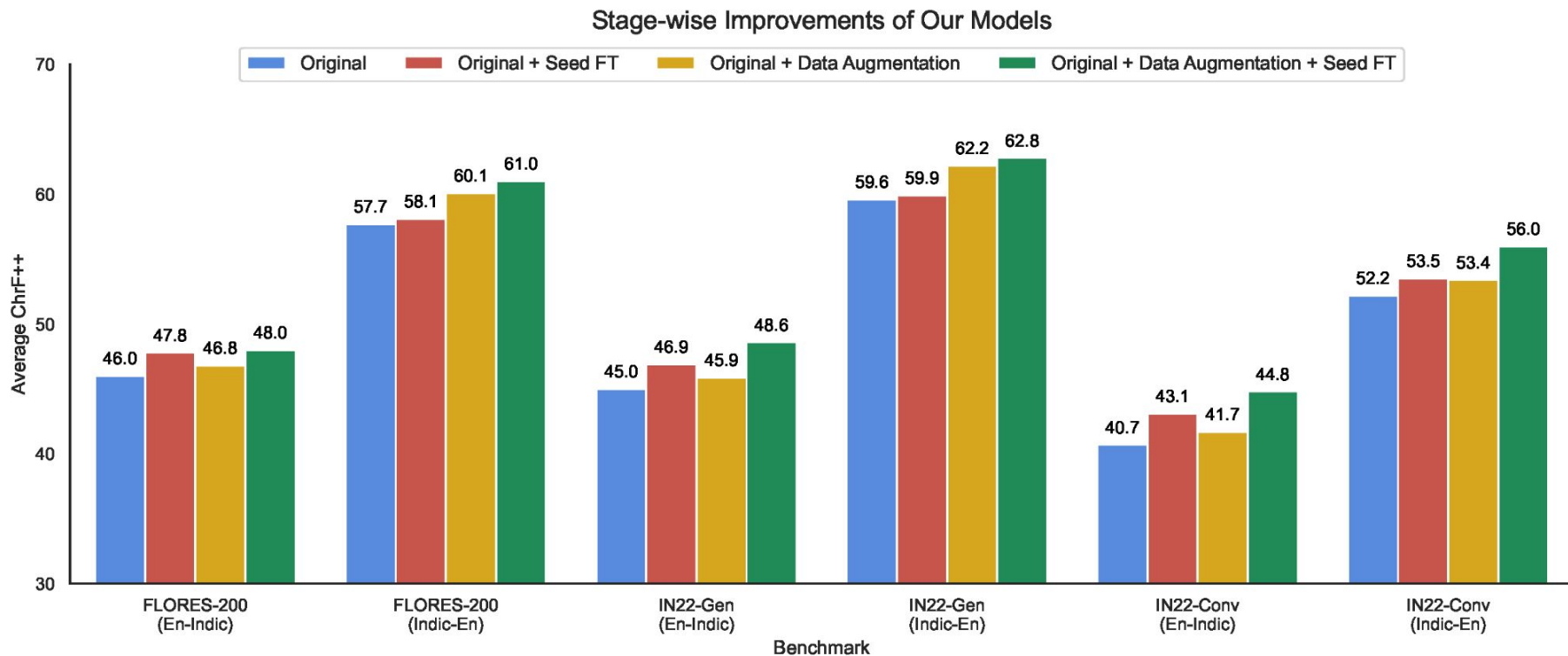


Our Training Strategy : Multi-stage Training

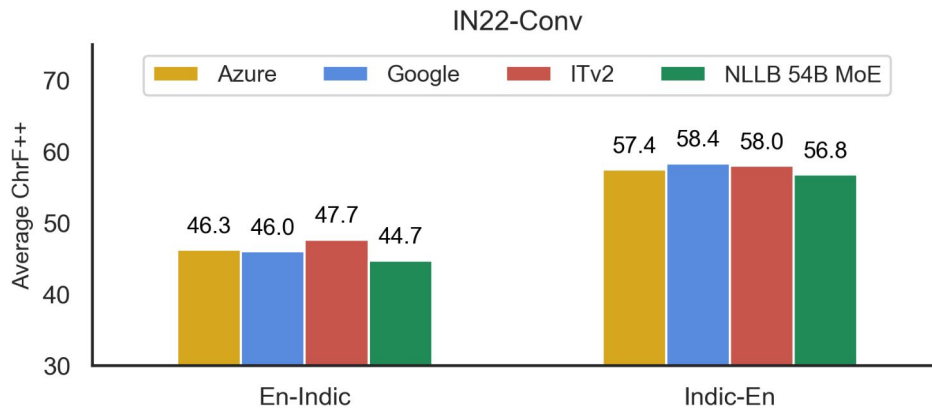
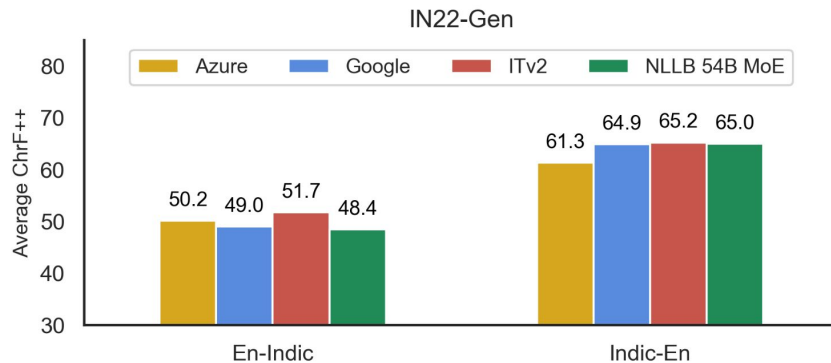
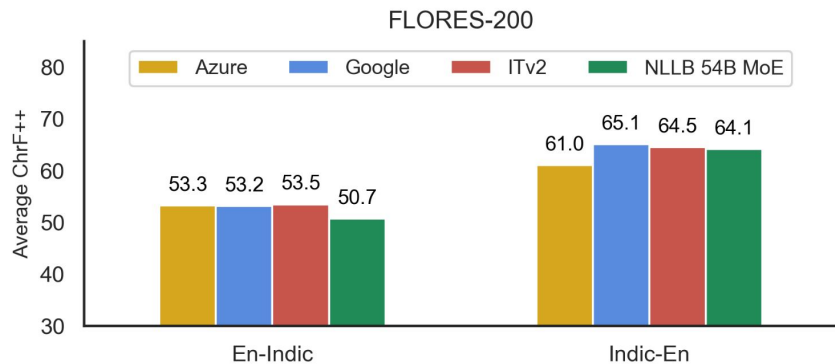


■ Train the Model from scratch ■ Finetune the model from stage 1

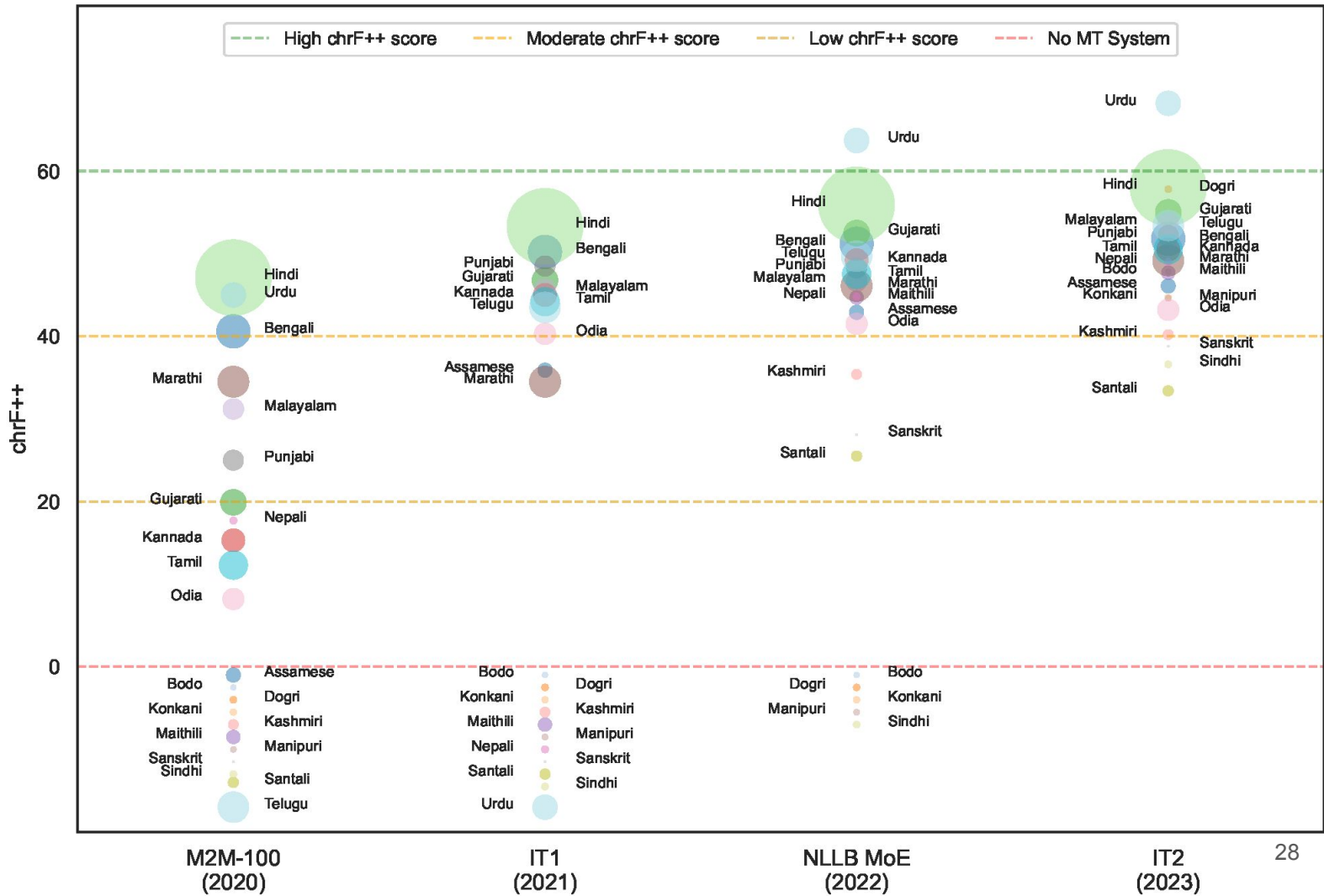
Stage-wise Improvement of Our Models



Summary: IndicTrans2 outperforms commercial systems on different benchmarks



Overall

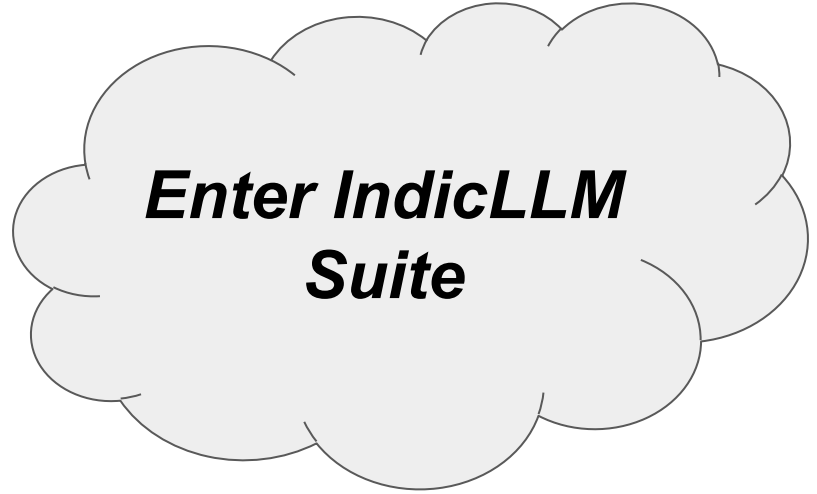
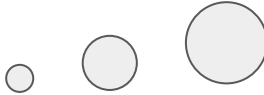


All That **Monolingual Data** Helped!

But how to get it?

When You Dont Have Data

- You mine it from the web
 - Ideal solution
 - Free of model biases
 - **Exhausting**
- You synthesize it
 - Non-ideal solution
 - Contains model biases
 - **Rapid**



IndicLLM Suite: Why is it even needed?

- ChatGPT does a good job in English
- Fails in Hindi (last time we checked)
- Fails in some culture specific questions, even in English
- *We want to capture what Indian people might ask! (Indic Original)*
- [Paper](#), [data](#)



You

How were Hanuman and Bhim related?



ChatGPT

Hanuman and Bhim are characters from Hindu mythology, specifically from the epic texts Ramayana and Mahabharata, respectively. They are not related in a familial sense but are often compared or referenced together due to their similar characteristics of immense strength and valor.

Hanuman is a central character in the Ramayana, where he is depicted as a vanara (monkey) who serves Lord Rama, the seventh avatar of the god Vishnu. He played a crucial role in the search for Rama's wife, Sita, who had been kidnapped by the demon king Ravana, and in the subsequent battle against Ravana's forces.

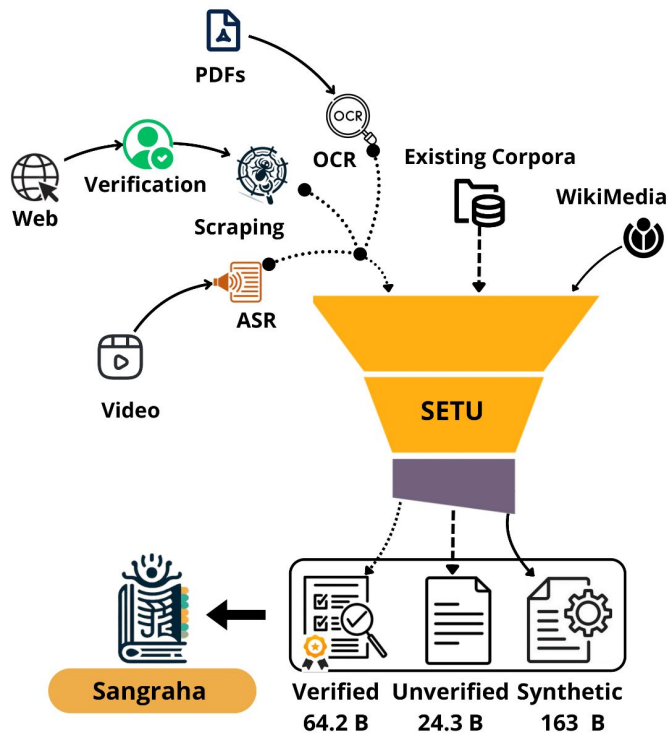
Bhim, on the other hand, is one of the Pandava brothers in the Mahabharata, which is another major Hindu epic. He is known for his exceptional strength and prowess in combat. Bhim's strength is often compared to that of Hanuman due to their similar feats of power.

While Hanuman and Bhim are not directly related in terms of family lineage, they are often mentioned together in discussions about strength, courage, and valor in Hindu mythology.

< 2 / 2 >    

Components

- Sangraha
 - Monolingual documents
 - Synthetic data
 - Verification
 - OCR
 - ASR
- IndicAlign
 - Instruct
 - Prompts by humans
 - Responses by LLMs
 - Toxicity data



IndicAlign

INSTRUCT

Component	Examples	Prompt source	Response source	Original / Translated
Indic ShareLlama	21.1 k	Human	Model	Translated
Dolly - T	15.0 k	Human	Human	Translated
OpenAssistant - T	19.9 k	Human	Human	Translated
WikiHow	26.4 k	Human	Human	Translated
IndoWordNet	74.2 M	Human	Human	Human
Anudesh	43.3 k	Human	Model	Translated
Wiki-Conv	144 k	Model	Model	Translated
Wiki-Chat	202 k	Model	Model	Translated

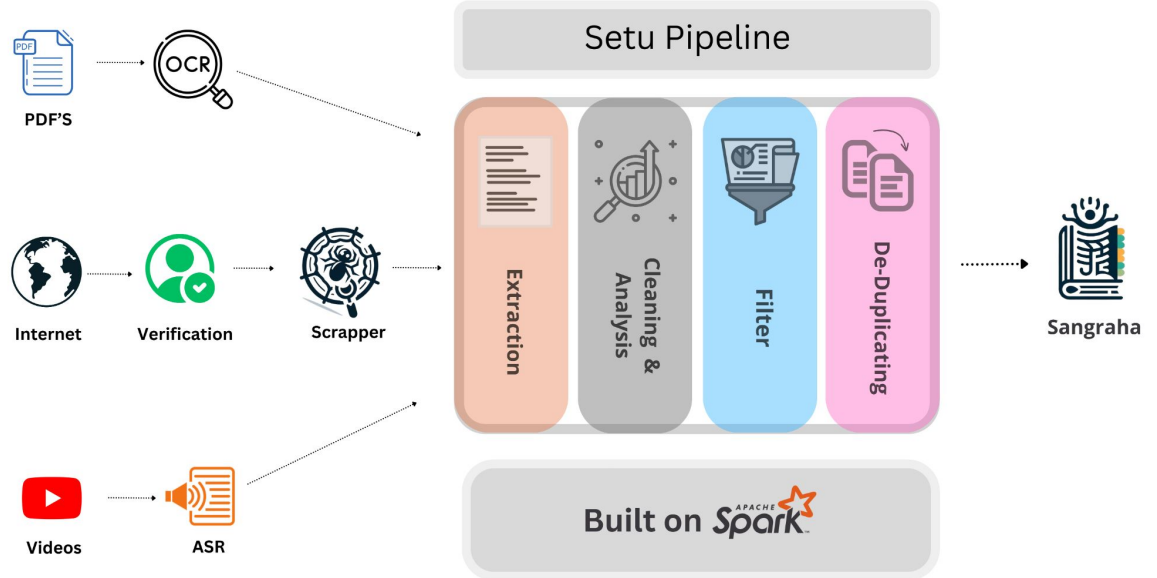
TOXIC

HH-RLHF	32.6 k	Human	Model	Translated
ToxicMatrix	90.3 k	Model	Model	Translated

Human
Model
Translated

Creating Sangraha

- OCR (**14.5B**)
 - PDFs from InternetArchive
 - \$\$\$
- Web scraping (**48B**)
 - Verified sources of data
 - Manual intervention for verification
- ASR (**1.2B**)
 - Lots of videos
 - HindiConformer model
- Setu
 - Standard cleaning process
 - Scalable
 - Priyam Mehta (AI Resident)



Sangraha Synthetic and Unverified

- MT and Xlit to the rescue
 - IndicTrans2 for bulk translation ([paper](#), [models](#))
 - IndicXLIT for bulk transliteration ([paper](#), [code & models](#))
 - English Wikimedia → 22 Indic languages
 - **90B tokens**
 - Translated content → Transliterate
 - **72B tokens**
 - Approach: split documents → translate sentences → assemble
 - Caveat: Document level phenomenon may be lost
 - *Note: Our work on synthetic data for LM training by [Doshi+, 2024](#)*
- MadLad and CulturaX
 - High quality data but unverified
 - **24B tokens**

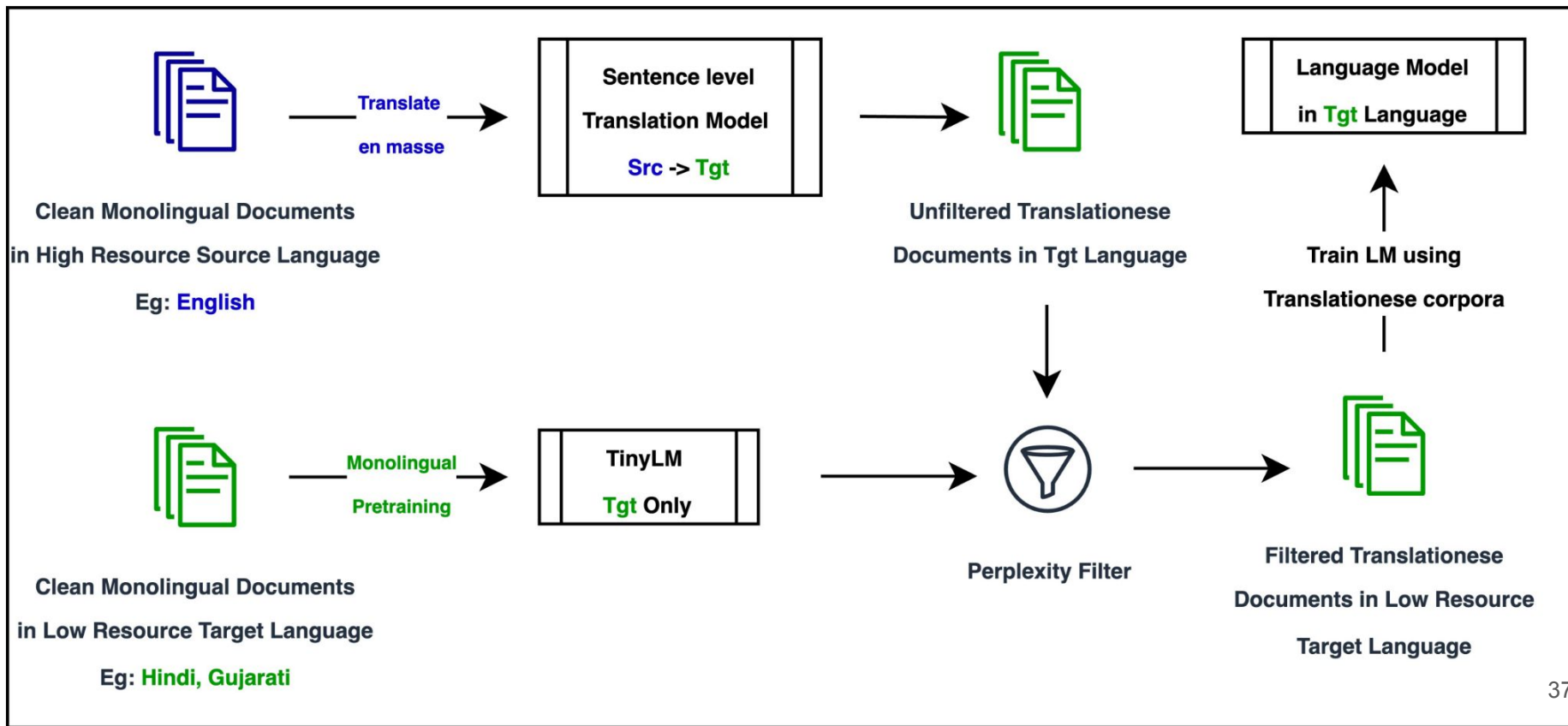
Speaking of Synthetic Data: How Good is It?

- Can synthetic data truly be useful for LLMs?
- Is cleaning needed?
- How do we verify impact?
- Enter **TinyLMs and Translationese** ([paper](#))
 - Work with **Meet Doshi** (CFILT, IIT Bombay)

Why Translation and TinyLMs?

- Translation
 - Cheap
 - Easy
 - Mass production
 - Somewhat reliable
- TinyLMs (and [babyllms](#))
 - Sandbox mode
 - Large scale experiments
 - Understand scaling laws
 - Low-compute requirement
 - 60-100M param models + 6B tokens < 1 day on a A100
 - [TinyLLAMA](#) library is fast
 - Our work: 28-85M param models (non-embedding) and ~6-10B tokens

Overview



Results

		Model	FLORES						
			EN-HI	HI-EN	Avg.	n.		Avg.	
M		BI-EN-HI-clean	46.56	51.7	49.13				
		BI-EN-HI_syn-parallel-filtered	44.12	50.64	47.38				
		BI-EN-HI_syn-nonparallel-filtered	45.65	51.29	48.47				
GU			EN-GU	GU-EN	Avg.	18	15.54	5	32.04
syn-GU_						56	14.39	4	31.14
syn-H	syn-GU_en-	BI-EN-GU-clean	26.44	35.3	30.87		-		-
sy	syn-GU	BI-EN-GU_syn-parallel-filtered	26.77	34.84	30.81	05	15.33	3	31.13
syn-l	syn-GU_ei	BI-EN-GU_syn-nonparallel-filtered	26.7	36.54	31.62		-		-

- Unfiltered translated data is *inferior*
- Filtering can **approach clean data performance**
- 10% clean data can help surpass full clean data performance

How About Bilingual and English-only models?

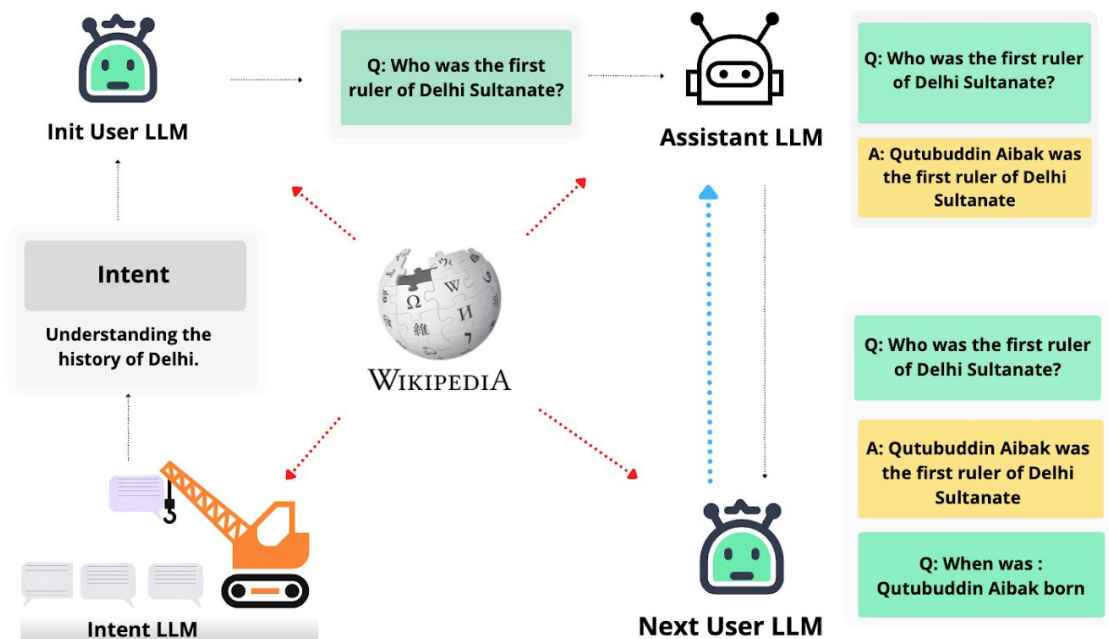
Model		sst2	cola	mrpc	qnli	qqp	rte	mnli-m	mnli-mm	stsb	Avg.
		acc						acc	pearson		
Mono	EN-clear	87.96							82.67		76.87
	syn-EN_hi-ur	87.16							78.9		71.84
	syn-EN_h	87.8							80.32		73.78
	syn-EN_h	87.1							81.1		74.60
	syn-EN_h	76.91							81.95		75.83
	BI-EN							77.27	82.07		75.88
	BI-EN-HI_syn-parallel-filtered	87.02	38.68	84.9	83.7	84.06	65.34	77.31	77.85	82.41	76.31
Bi	BI-EN-HI_syn-nonparallel-filtered	89.79	38.68	84.9	83.7	84.06	65.34	77.15	77.55	83.01	76.40
	BI-EN_syn-HI_syn-filtered	87.95	30.05	84.9	83.7	83.97	63.89	75.63	76.24	82.24	74.29
	BI-EN_syn-HI_syn-filtered + 10%	89.1	35.45	85.34	84.53	84.18	65.7	76.64	77.24	82.1	75.59

Era of the **synth** is upon us!

- Synthetic English works too!
- Clean for one and synthetic for another also works!

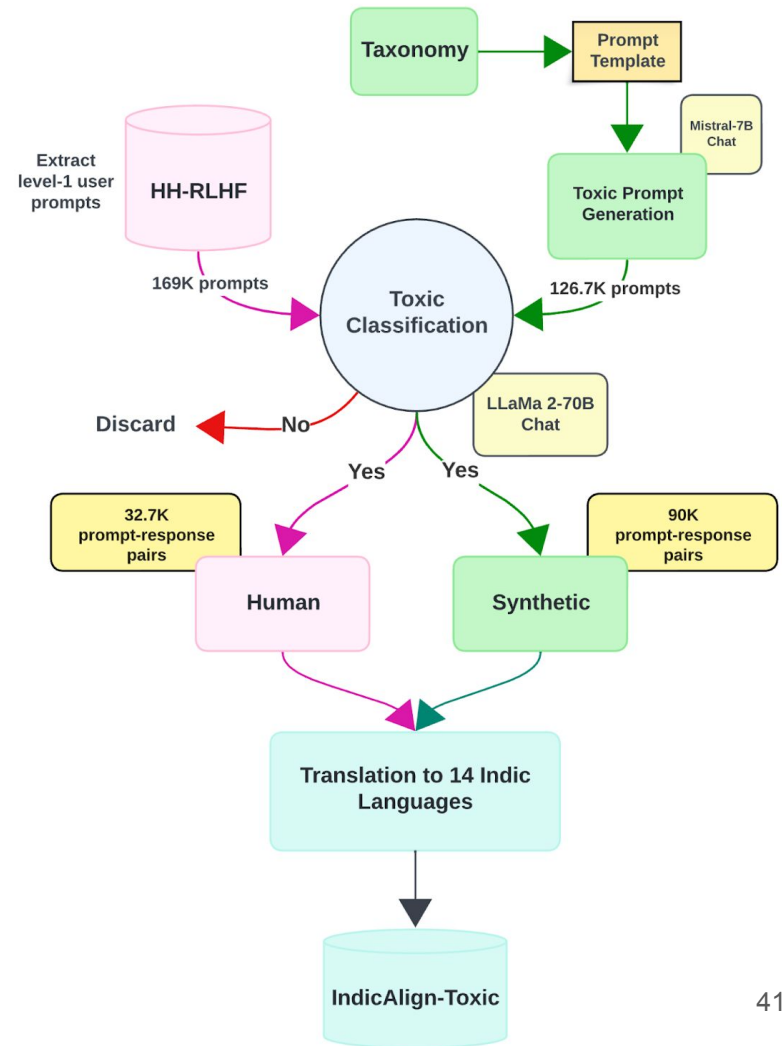
Time for Alignment: Instruction Data

- Leverage [IndoWordnet](#)
 - Gloss, synonyms, cross-linkages
 - Basic but useful
 - 18 languages, **74M pairs**
- Leverage instruction datasets
 - Dolly, OpenAssistant, and WikiHow
 - 14 languages, **~80K pairs**
- LLM generated
 - Context grounded QA pairs
- Human in the loop
 - **43K conversations**
 - Translate-test
- Translate and Transliterate



Time for Alignment: Toxicity Data

- Taxonomy by **Safi Khan (AI Resident)**
 - Used to construct prompt templates
 - “Unsafe” model to create prompts
 - Mistral 7B chat works
- Classification
 - Toxic Part of HH-RLHF
 - Toxic part of previous step
 - LLAMA2-7B Chat as judge
 - 32.7K from HH-RLHF
 - 90K from Taxonomy+Prompts+Mistral
- Translate and transliterate it all



How Good Are LLMs for Indic MT?

Background: LLM Prompting for MT

Retrieve example for given input → Condition → Elicit response

Example from the
database as
context



Hindi Sentence: "यह अभी भी शुरुआती दिनों में है।"
English Sentence: "Its still in the early stages."
###

Input
Source



Hindi Sentence: "डॉ. एहुड उर, नोवा स्कोटिया के हैलिफैक्स में डलहौज़ी
विश्वविद्यालय ... "

LLM

Obtain
Machine
Translation

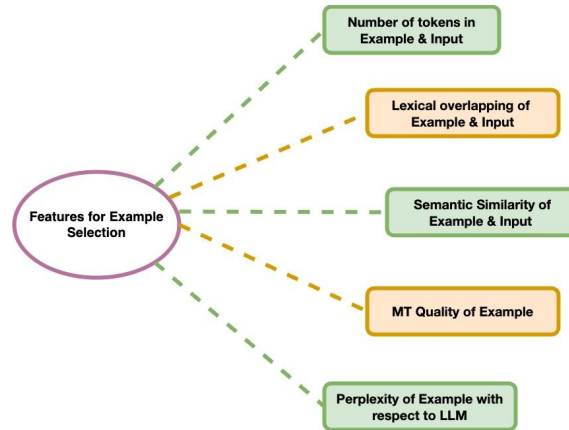
Obtain feature metrics

LaBSE-InSrc	Cmt-InSrc	Cmt-SrcTgt	NumTokTgt
LaBSE-InTgt	Cmt-InTgt	NumTokIn	PPL-SrcTgt
chrF-InSrc	LaBSE-SrcTgt	NumTokSrc	PPL-SrcTgtIn

Compute Translation
score
(COMET/BLEU score)

CTQ-Scorer: Combine Features For Re-ranking Examples

- In-context learning relies on various features of selected examples, including their quality and relevance.
- Existing works typically concentrate on individual features in isolation.
- Several features correlate with prompting performance, the correlation is weak (Zhang et al., 2023).



In this work, we propose a general framework for combining different features influencing example selection. ([Kumar+, 2023](#))

CTQ-Scorer

BM25

Predict Comet
Score

Algorithm 1 Algorithm for creation of data to train the CTQ Scorer regression model

- 1: **Inputs** Held-out example pairs (x, y) , example database D
- 2: **Outputs** Training data for CTQ Scorer regression model
- 3: **procedure** CREATETRAININGDATA
- 4: **for** a given (x, y) from held-out example pairs **do**
- 5: Perform *Candidate Shortlisting* and retrieve K candidate examples from D
- 6: Each of the tuple (x_p, y_p) in K candidate examples is a prompt candidate
- 7: **for** a given (x_p, y_p) **do**
- 8: Generate the 1-shot translation y' of x using (x_p, y_p) as prompt example
- 9: Generate the Translation score using any sentence-level MT metric, $\text{xlate_score}(x, y', y)$
- 10: $\text{ctq} = \text{xlate_score}(x, y', y)$
- 11: $\text{featset}(x_p, y_p, x) = \text{Feature Extraction}$ using the triple (x_p, y_p, x)
- 12: $\text{Training Instance} = \text{featset}(x_p, y_p, x), \text{ctq}$
- 13: **return** *All Training Instances*

overiap based sentence
retriever

Translation

Results (4-shot)

Selection Method	bn	gu	hi	de	fr	ru	Average
Random Selection	40.07	38.27	44.52	63.05	*70.89	*49.40	51.03
BM25	38.93	38.42	45.18	62.14	*70.82	45.76	50.21
R-BM25	39.97	38.16	45.20	62.94	*70.31	*49.28	50.98
CTQ (<i>ours</i>)	42.99	41.77	50.03	64.77	71.28	50.85	53.62
CTQ-QE (<i>ours</i>)	38.56	40.45	45.40	64.13	71.33	50.72	51.76

COMET scores for translation into English using different example selection methods.

The highest scores are in **bold** text.

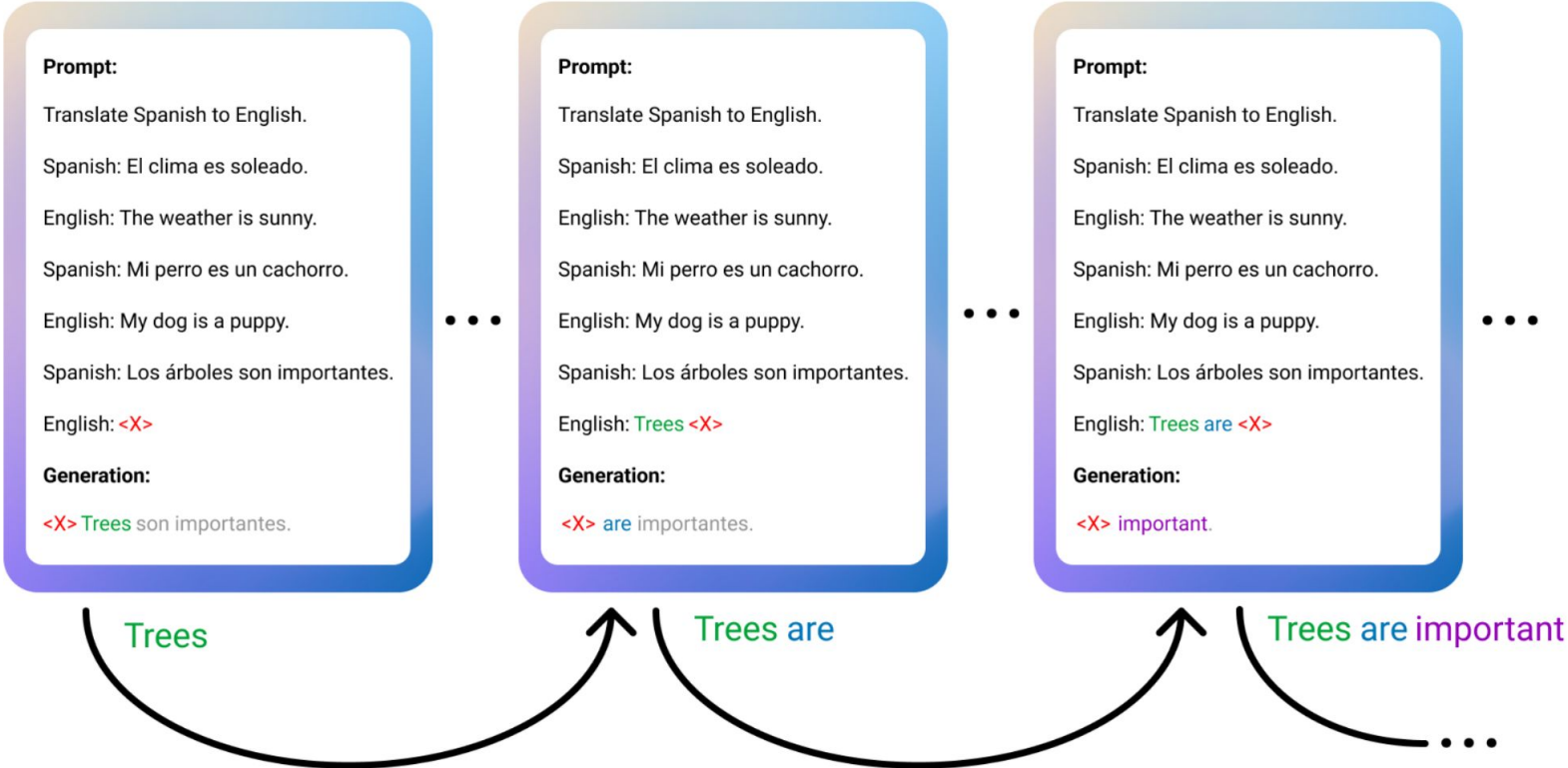
We compared CTQ with Random, BM25 and R-BM25 for statistical significance.

All comparisons with CTQ are statistically significant ($p < 0.05$) (except results marked with *) as per paired bootstrap sampling.

Example selection using the CTQ Scorer outperforms other methods (2.5 - 4.5 COMET points).

Can We Use **Encoder-Decoder PLMs** For **MT**?

Enter SAP with mT5 (no, not the company; [Patel+ 2023](#))

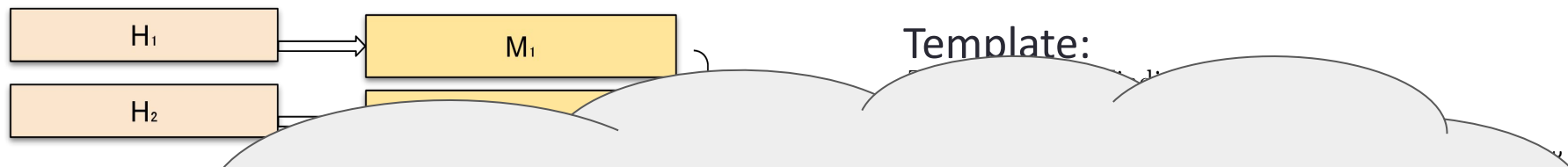


Leveraging Relatedness ([Puduppully+ 2023](#))

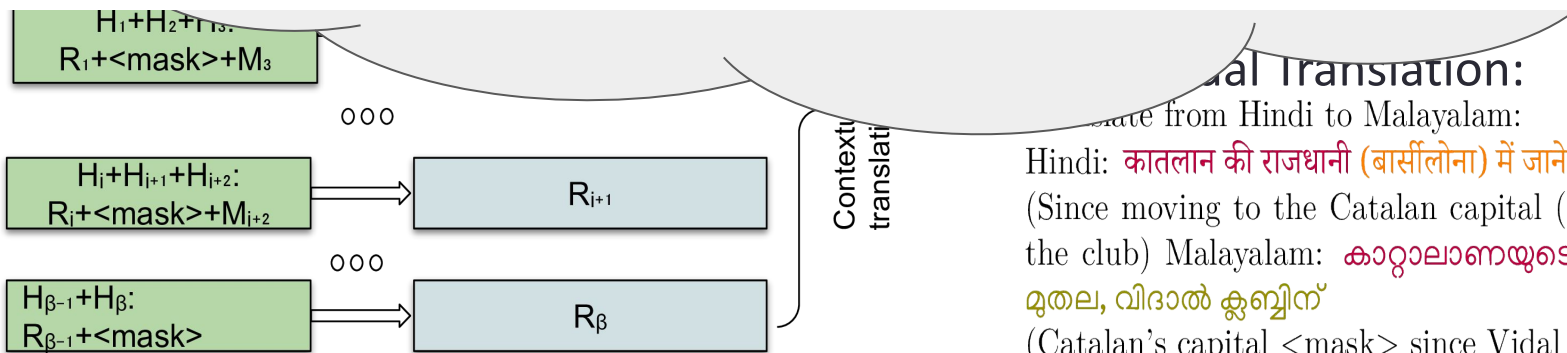
	Boy Subject	apple Object	eats Verb
hin:	लड़का (ladka)	सेब (seb)	खाता है (khaata hai)
mal:	ആൺകുട്ടി (aankutti)	ആപ്പിൾ (aappill)	കഴിക്കുന്നു (kazhikyunnu)

Monotonicity is often overlooked

DecoMT: Prompt mT5 Independently (5), then Contextually (10)



$$\begin{aligned}
 [\text{psrc}] : X'_1 \quad [\text{ptgt}] : Y'_1 \dots [\text{psrc}] : X'_K \\
 [\text{ptgt}] : Y'_K \quad [\text{src}] : X \quad [\text{tgt}] : \quad (2)
 \end{aligned}$$

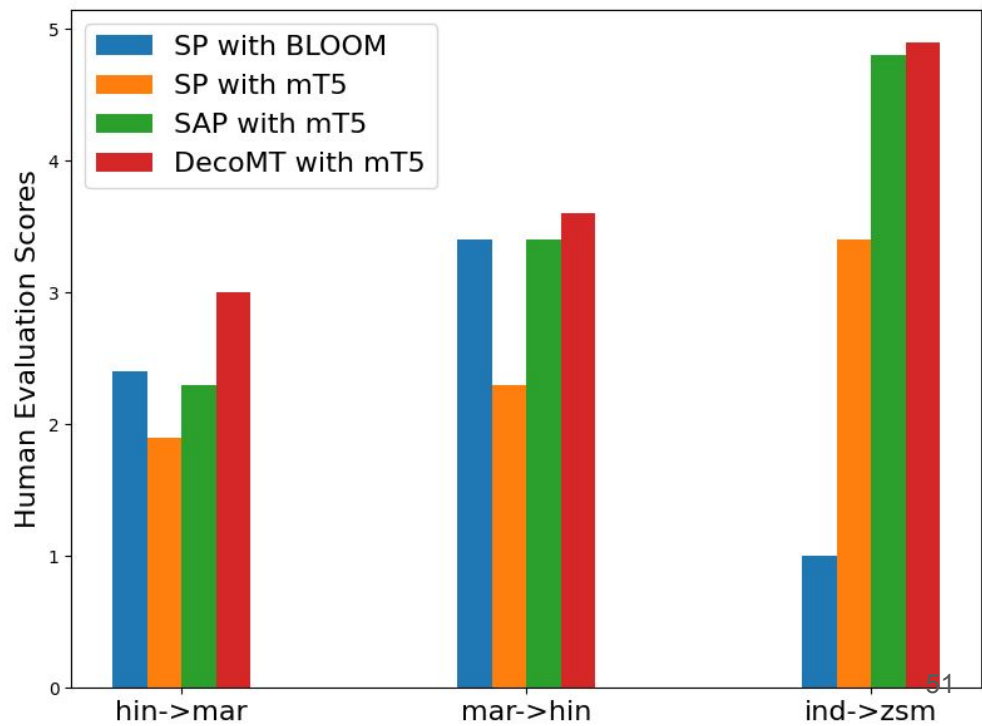
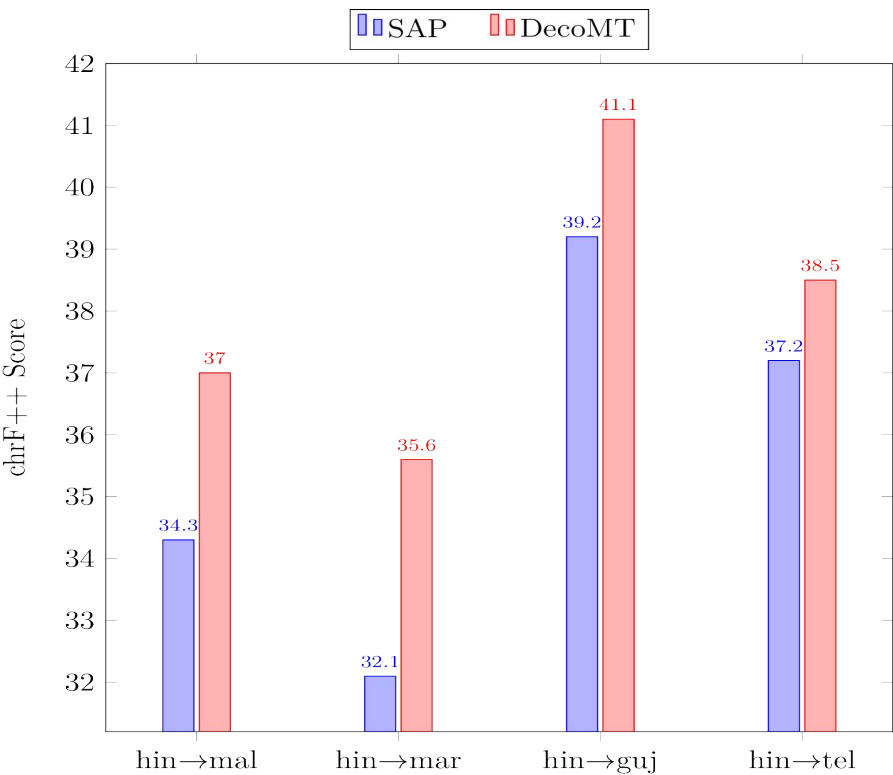


Contextual Translation:

Translate from Hindi to Malayalam:

Hindi: कातलान की राजधानी (बासीलोना) में जाने के बाद से, विडाल ने क्लब के (Since moving to the Catalan capital (Barcelona), Vidal has for the club) Malayalam: കാറ്റാലാണയുടെ തലസ്ഥാനമായ <mask> മുതല, വിദാൽ ക്ലബ്ബിന് (Catalan's capital <mask> since Vidal for club)

Results: Monotonically Translated Chunks Help!

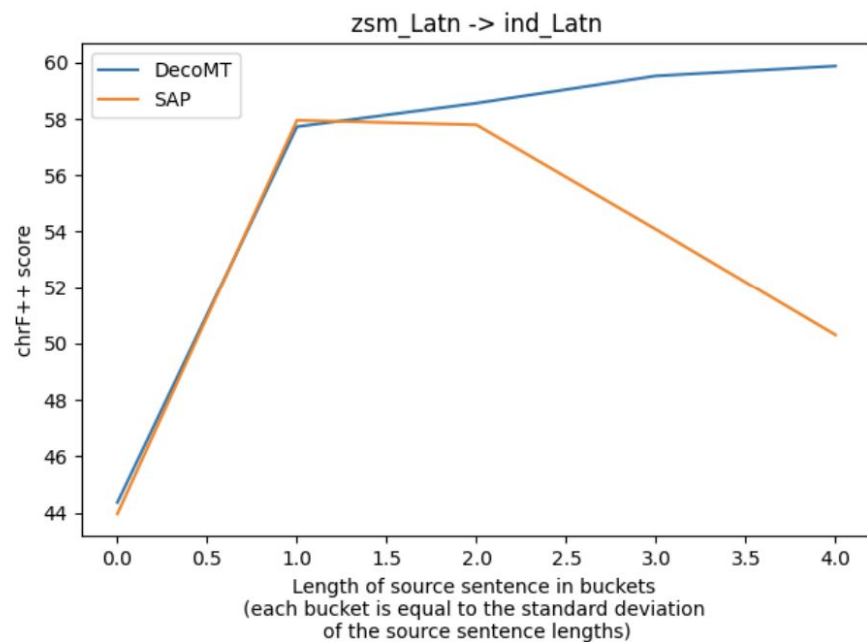
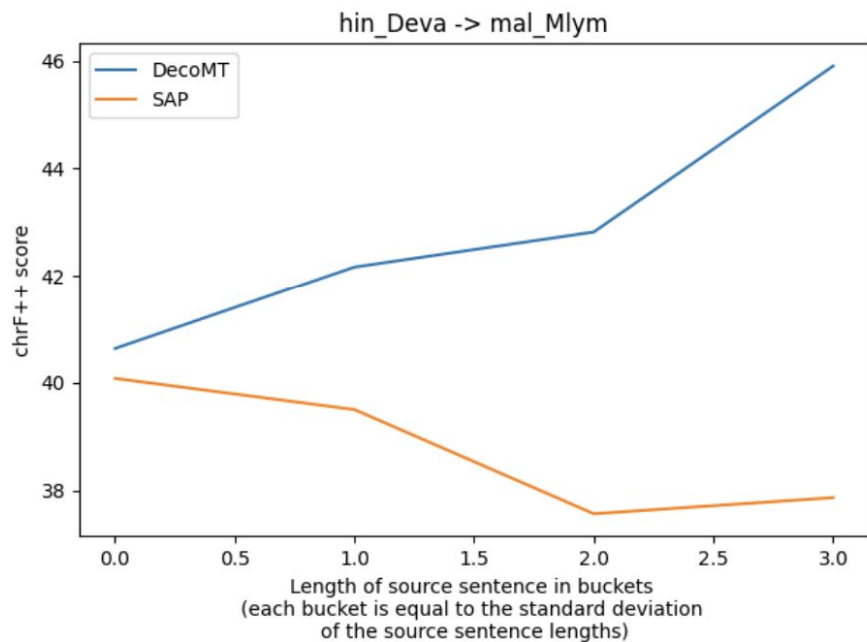


Elimination of Off-Target Translations

- SP: Predict a phrase incrementally
- SAP: Predict a phrase incrementally
- Massive reduction in off-target translations
- Big problem that plagues non-English centric MT

	SP		SAP		DecoMT
	BLOOM	XGLM	mT5	mT5	mT5
hin→mal	23.6	100.0	14.4	0.4	0.0
mal→hin	8.4	0.0	4.4	1.4	0.2
hin→mar	21.2	96.3	35.2	10.0	0.8
mar→hin	1.3	20.0	2.6	1.1	0.2
hin→guj	10.2	99.7	3.8	0.2	0.0
guj→hin	3.3	0.0	1.9	0.4	0.2
zsm→ind	–	48.8	23.3	17.7	13.1
ind→zsm	–	94.2	59.7	47.3	30.1
rus→ukr	–	84.3	1.7	0.2	0.0
ukr→rus	–	0.6	0.5	0.1	0.0
spa→por	0.2	0.4	3.4	0.9	0.2
por→spa	0.0	0.5	0.6	0.3	0.1

Robustness to Longer Sequences



But What If An LLM
Did Not Support Indic Languages?

Llama :

॑



॑

ikta hai

RomanSetu

Romanization as a bridge

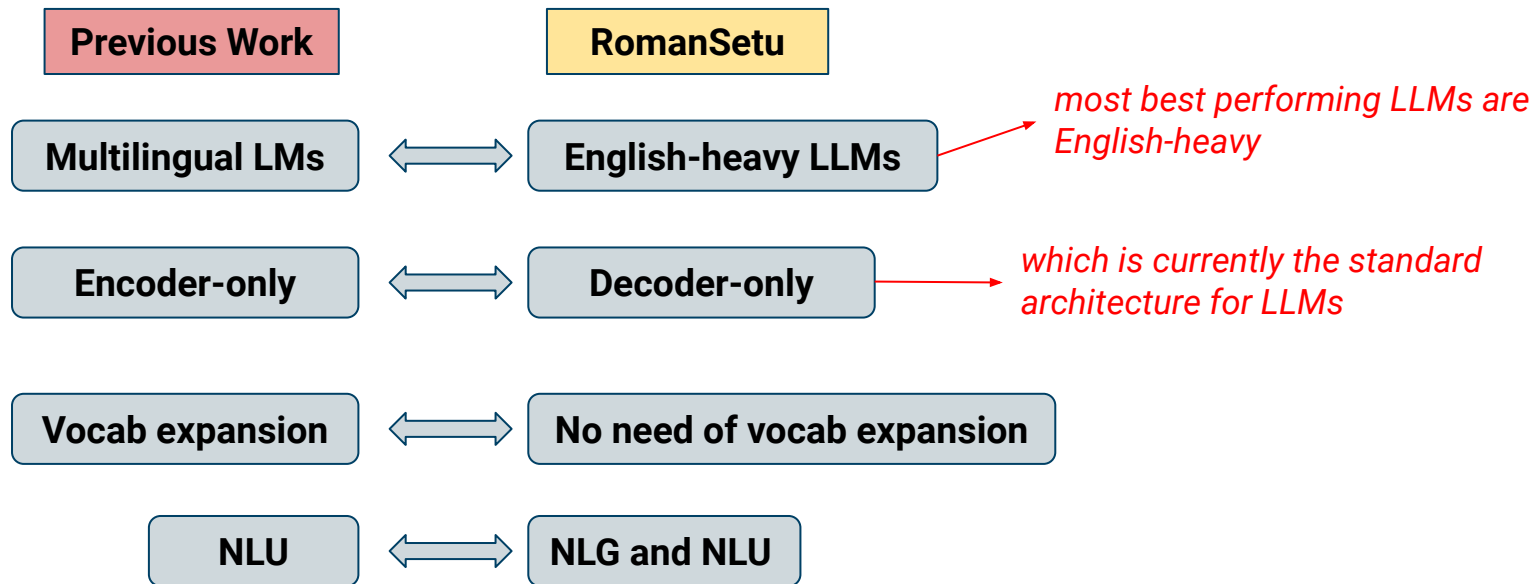
Why Romanization?

- Significant part of web data is code mixed and Romanized
- Major LLMs have seen such data
 - Chance to transfer from English
- Byte based BPE oversegments
- But, Romanization doesn't

Language	N	R	R-IndicNLP
Guj			16
			3
			5.56
Marathi	8.7	5.64	4.84
Tamil	12.11	4.89	5.35

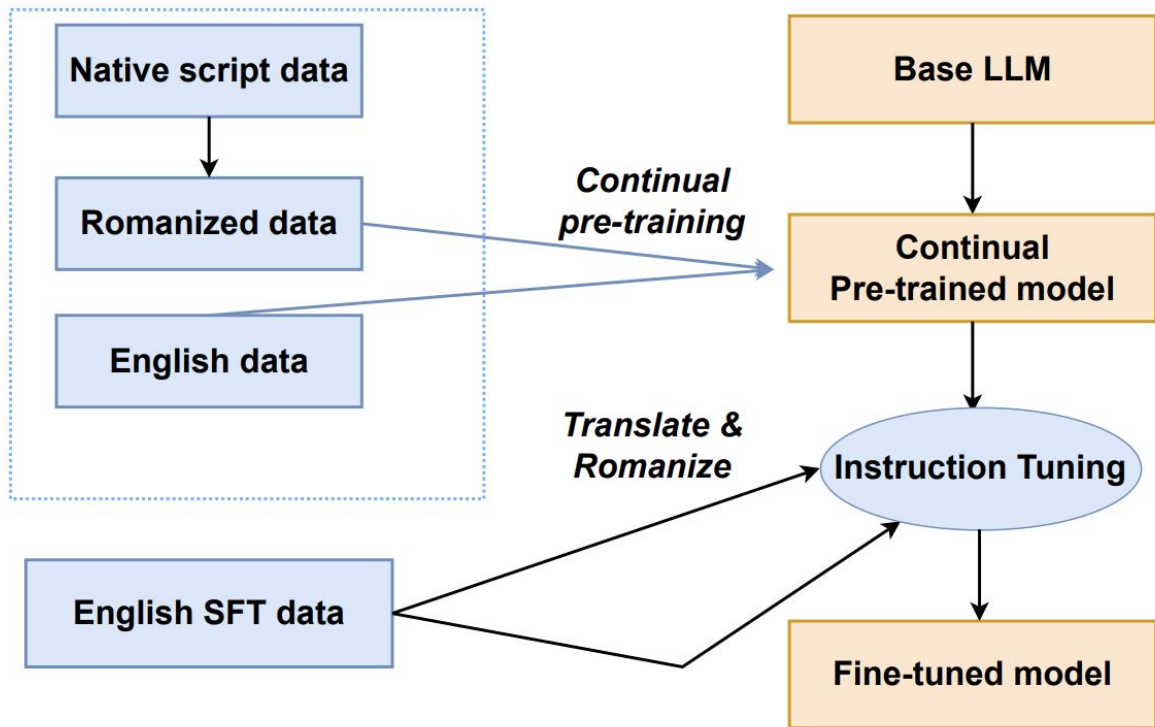
A chance for efficiency?

Whats Different About RomanSETU?



Creating the bridge

- Continual pre-train for a bit
 - Romanization (R)
 - Native (N)
 - **Both!**
- SFT
 - Again N, R, or N+R
- Ideally
 - $R > N$
- Realistically
 - R approaches N



How Does LLaMa2 See Romanization?

- Sequence vector as **last token representation of last layer**
- Compare English, Native, Romanized (trilingual)
- *Higher overlap between English and Romanized representations*

Language	E - N	E - R	N - R
Gujarati	0.39	0.47	0.51
Hindi	0.40	0.50	0.34
Malayalam	0.40	0.46	0.52
Marathi	0.44	0.48	0.58
Tamil	0.44	0.43	0.53

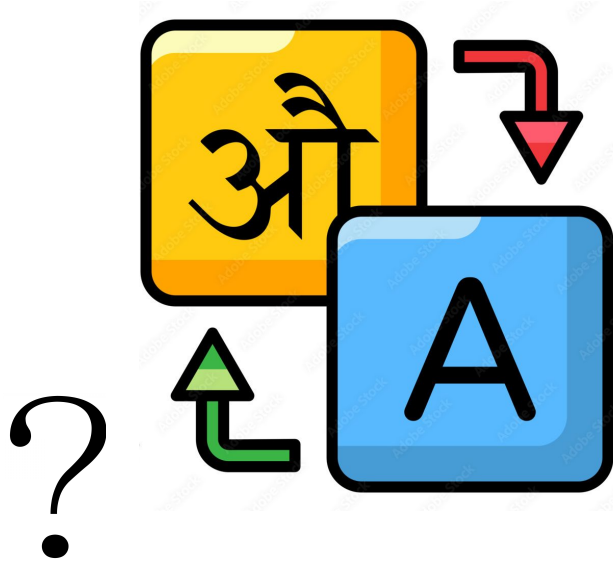
Impressive Performance

- Significantly **beats** native script performance
- Better **cost efficiency** for same number of tokens
- *Large margin* over native script for generation in Indic languages
- Mapping back to original script may not indicate true gains
 - **Transliteration errors** to blame
- Future work
 - *Devanagari mapped data* for fine-tuning recent LLMs?
 - Evaluation?

	Script	BaseLLM		CPT		IFT
		1-shot	3-shot	1-shot	3-shot	0-shot
XX-En translation	N	22.46	23.42	38.54	37.07	49.78
	R	22.52	23.52	42.53	41.64	50.75
En-XX translation	N	13.95	14.25	25.55	26.19	37.40
	R	14.20	12.02	29.55	30.77	46.87
XLSum	N	6.88	-	7.59	-	7.77
	R	10.16	-	12.44	-	12.56
IndicHeadline	N	13.66	-	18.04	-	12.61
	R	15.56	-	18.92	-	16.03

Towards Evaluation

Which Metric Is Reliable?

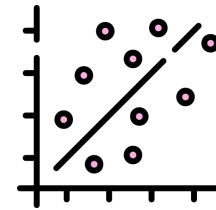


Evaluation of translations from English to other languages is under-studied

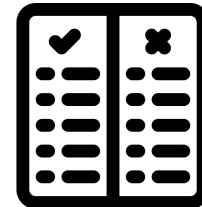
Several evaluation metrics proposed and studied for to-English translation.

- BLEU
- METEOR
- TER
- Embedding Averaging
- BERTScore
- PRISM
- BLEURT...
-

Meta Evaluation of Metrics



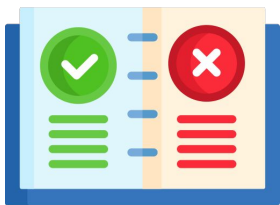
Correlations with human scores



Advantages and shortcomings

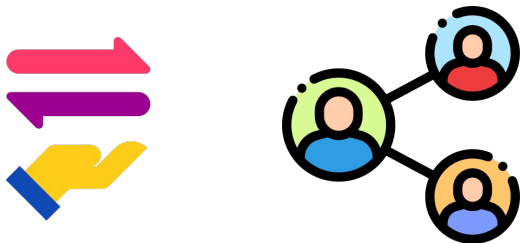
Automatic Evaluation of Machine Translation

It is important to study evaluation metrics for other languages instead of naively adopting the metrics proposed for English



Their own Grammar rules

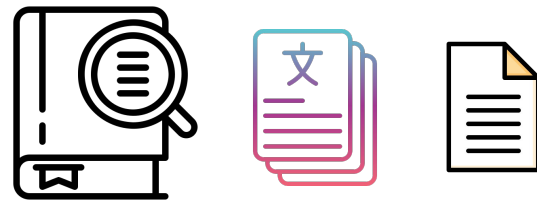
Shared and Borrowed vocab



Subject	Verb	Object
I	went	shopping

Subject	Object	Verb
நான்	கடைக்கு	போனேன்
(I)	(shopping)	(went)

Different sentence structure



Varying amounts of resources

Evaluation of Translation to Indian languages

In this work we focus on Indian languages.

5 languages belonging to 2 different language families

Dravidian languages

தமிழ்

- Tamil

മലയാളം

- Malayalam

Indo Aryan languages

- Hindi हिन्दी

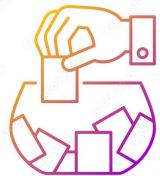
- Marathi मराठी

- Gujarati ગુજરાતી



Flores dataset

Collecting data



Select 200 random sentences



<Source sentence>

- mBART
- Bing API
- Google API
- CVIT-IIITH
- IndicTrans
- mT5
- NLLB



<Translation outputs>



.

.

.

.

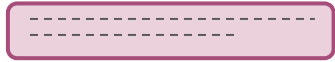
.

.

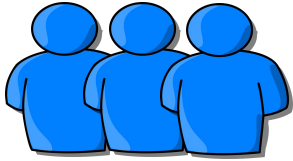


Total: 7000 samples

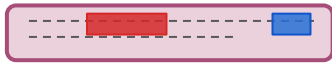
Collecting Human Annotations - MQM framework



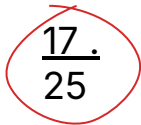
Output of system k for source sentence s



Bilingual Expert annotators

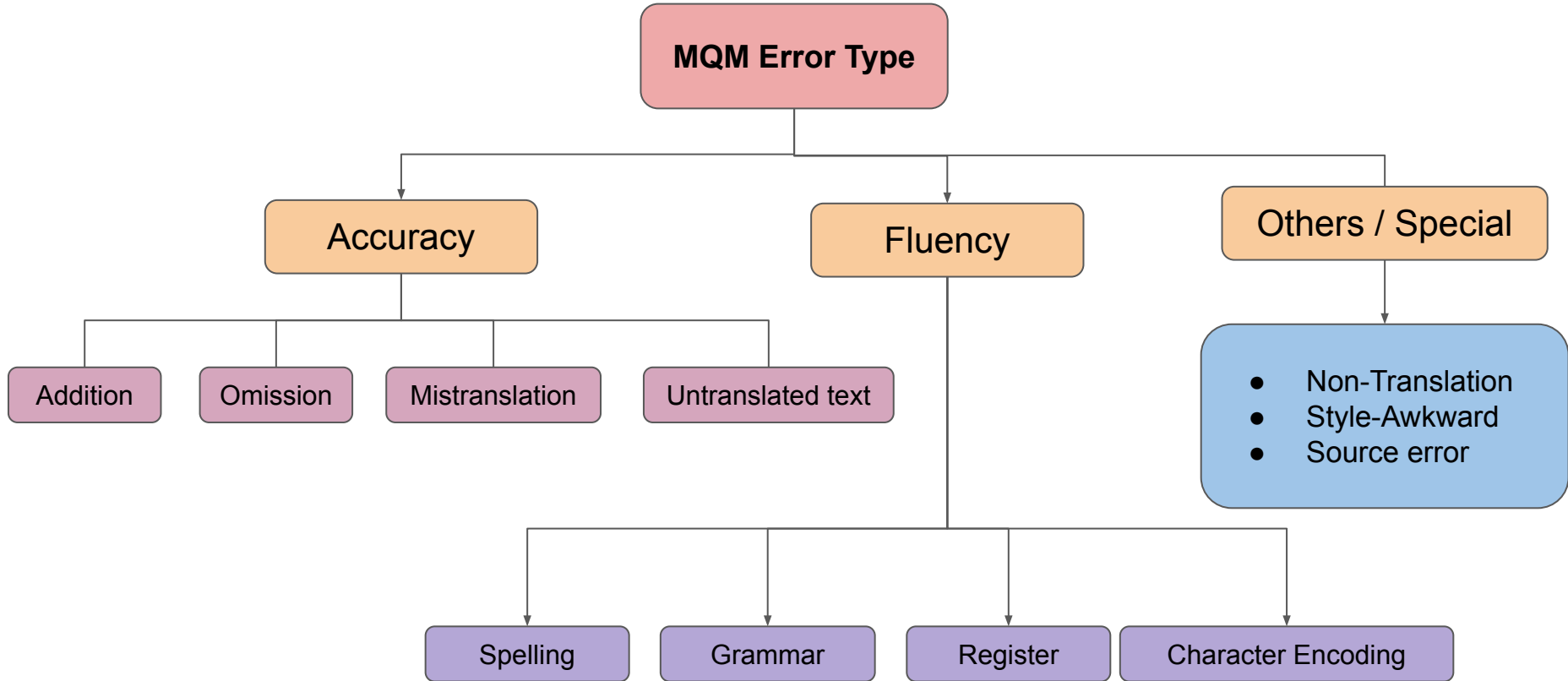


Highlight any minor / major errors in the text and judge the output along multiple criteria



Also provide an overall score

MQM framework - Error categories



Example Annotation

Source It was one of the major stops during Henry Louis Gates' PBS special Wonders of the African World.

Google API ஹென்றி லூயிஸ் கேட்ஸின் ஆப்பிரிக்க உலகின் பிபிஎஸ் சிறப்பு அதிசயங்களின் போது இது ஒரு முக்கிய நிறுத்தமாக இருந்தது.

Annotations **Fluency Spelling**: Low
Fluency Grammar : Very High
Fluency Grammar : High

Error Statistics of each system



Average computed human scores for each system							
lang	IndicTrans	Bing API	CVIT-IIITH	Google API	mBART	mT5	NLLB
gu	22.639	23.179	19.034	21.686	0.000	20.067	22.490
hi	20.120	14.405	14.962	19.484	15.703	18.012	18.445
mr	18.484	17.934	17.586	15.750	5.773	14.441	18.618
ml	22.676	22.617	17.844	21.955	17.355	20.169	21.515
ta	17.978	16.516	11.933	16.651	13.522	15.994	17.578
avg	20.379	18.930	16.272	19.105	10.471	17.737	19.729

Ranking of the systems based on expert human scores

Correlations of various metrics with human scores ([Sai+, 2023](#))

Metric	gu		hi		mr		ml		ta		Average	
	ρ	τ	ρ	τ	ρ	τ	ρ	τ	ρ	τ	ρ	τ
BLEU 1	0.364	0.255	0.266	0.187	0.228	0.148	0.393	0.331	0.316	0.213	0.314	0.227
BLEU 2	0.329	0.247	0.280	0.192	0.190	0.135	0.331	0.302	0.291	0.205	0.284	0.216
BLEU 3	0.294	0.234	0.265	0.186	0.134	0.119	0.250	0.271	0.227	0.182	0.234	0.198
BLEU 4	0.235	0.215	0.245	0.171	0.091	0.103	0.180	0.246	0.171	0.168	0.184	0.181
SacreBLEU	0.293	0.239	0.255	0.168	0.164	0.132	0.274	0.298	0.244	0.189	0.246	0.205
ROUGE-L	0.350	0.251	0.295	0.204	0.206	0.132	0.376	0.322	0.308	0.206	0.307	0.223
chrF++	0.408	0.287	0.299	0.205	0.260	0.170	0.411	0.338	0.361	0.250	0.348	0.250
TER	0.304	0.237	0.263	0.196	0.203	0.135	0.343	0.307	0.272	0.199	0.277	0.215
EA	0.331	0.181	0.086	0.066	0.143	0.054	0.397	0.301	0.203	0.149	0.232	0.150
VE	0.380	0.265	0.274	0.183	0.234	0.153	0.412	0.331	0.337	0.227	0.327	0.232
GM	0.394	0.266	0.234	0.162	0.241	0.147	0.426	0.338	0.382	0.264	0.335	0.235
LASER embs	0.094	0.156	0.135	0.123	0.159	0.069	0.357	0.295	0.126	0.099	0.174	0.148
LabSE embs	0.504	0.319	0.149	0.185	0.319	0.204	0.416	0.337	0.339	0.286	0.345	0.266
mBERT	0.448	0.297	0.337	0.231	0.301	0.194	0.462	0.367	0.413	0.281	0.392	0.274
distilmBERT	0.431	0.289	0.316	0.220	0.281	0.181	0.465	0.371	0.415	0.278	0.382	0.268
IndicBERT	0.456	0.308	0.346	0.235	0.281	0.182	0.440	0.357	0.402	0.282	0.385	0.273
MuRIL	0.465	0.322	0.353	0.243	0.292	0.184	0.449	0.369	0.410	0.290	0.394	0.282
PRISM	0.114	0.024	0.178	0.124	0.131	0.084	0.089	0.064	-0.040	-0.040	0.094	0.051
BLEURT-20	0.509	0.371	0.296	0.300	0.409	0.286	0.496	0.390	0.491	0.374	0.440	0.344
COMET-QE-DA	0.417	0.324	0.535	0.404	0.551	0.430	0.386	0.341	0.531	0.391	0.414	0.378
COMET-QE-MQM	0.387	0.309	0.590	0.403	0.577	0.392	0.438	0.392	0.571	0.399	0.513	0.379
COMET-DA	0.557	0.403	0.581	0.390	0.426	0.306	0.531	0.419	0.529	0.412	0.525	0.386
COMET-MQM	0.465	0.360	0.529	0.370	0.686	0.459	0.508	0.392	0.597	0.432	0.557	0.402

Indic COMET

- IndicCOMET - We finetune COMET metric variants using the MQM annotations

Metrics	gu		hi		mr		ml		ta		Avg.	
	ρ	τ	ρ	τ	ρ	τ	ρ	τ	ρ	τ	ρ	τ
COMET-DA	0.487	0.359	0.380	0.319	0.422	0.302	0.529	0.421	0.525	0.410	0.469	0.362
COMET-MQM	0.422	0.346	0.528	0.370	0.455	0.314	0.493	0.380	0.588	0.429	0.497	0.367
IndicCOMET _{XLM}	0.437	0.353	0.609	0.397	0.413	0.311	0.559	0.418	0.585	0.426	0.521	0.381
IndicCOMET _{DA}	0.431	0.339	0.554	0.384	0.436	0.310	0.526	0.410	0.587	0.433	0.507	0.375
IndicCOMET _{MQM}	0.446	0.360	0.616	0.419	0.463	0.331	0.566	0.416	0.597	0.441	0.537	0.393

Indic COMET - Zero-shot performance

Metrics	gu	hi	mr	ml	ta
COMET _{DA}	0.359	0.319	0.302	0.421	0.410
COMET _{MQM}	0.346	0.370	0.314	0.380	0.429
IndicCOMET _{MQM}	0.355	0.395	0.322	0.394	0.430

On the ACES Translation Accuracy Challenge Set, we evaluate robustness scores as follows:

- IndicCOMET_{MQM} = 0.306
- COMET_{MQM} = 0.272

Digging Deeper Into **Zero-Shot** Settings!

Underresourced languages

- Work by [Singh+, 2024](#)
- Use existing models on underresourced Indian languages
- Trends are similar to higher resourced languages
- Comet performed well as usual
- GEMBA (GPT-4) did badly

Metric	Assamese		Maithili		Kannada		Punjabi		Average	
	τ	ρ	τ	ρ	τ	ρ	τ	ρ	τ	ρ
BLEU 1	0.063	0.072	-0.131	-0.047	-0.017	-0.046	-0.002	-0.162	-0.022	-0.046
BLEU 2	0.058	0.081	0.078	-0.028	0.016	0.035	-0.016	0.065	0.034	0.038
BLEU 3	0.020	0.036	-0.028	-0.072	0.111	0.061	-0.055	0.023	0.012	0.012
BLEU 4	0.001	0.026	-0.032	-0.036	-0.088	-0.110	-0.023	0.065	-0.036	-0.014
SacreBLEU	0.075	0.104	0.199	0.265	0.103	0.155	0.098	0.154	0.119	0.170
ROUGE-L	0.088	0.128	0.052	0.055	0.005	0.003	-0.074	0.065	0.018	0.063
chrF++	0.160	0.254	0.252	0.366	0.145	0.228	0.164	0.255	0.180	0.276
TER	0.123	0.158	0.257	0.403	0.131	0.199	0.17	0.24	0.17	0.25
LASER embs	0.097	0.191	0.119	0.306	0.139	0.275	0.036	0.042	0.098	0.204
LabSE embs	0.128	0.194	0.125	0.169	0.219	0.366	0.19	0.303	0.166	0.258
mBERT	0.131	0.247	0.212	0.388	0.165	0.248	0.234	0.281	0.186	0.291
distilmBERT	0.139	0.267	0.25	0.416	0.169	0.263	0.245	0.306	0.201	0.313
IndicBERT	0.199	0.29	0.235	0.389	0.191	0.276	0.237	0.311	0.216	0.317
MuRIL	0.206	0.324	0.309	0.476	0.162	0.239	0.204	0.269	0.220	0.327
BLEURT-20	0.119	0.185	0.32	0.44	0.279	0.488	0.28	0.352	0.250	0.366
COMET-DA	0.228	0.298	0.172	0.264	0.281	0.390	0.300	0.358	0.245	0.328
COMET-MQM	0.26	0.381	0.199	0.291	0.290	0.410	0.266	0.334	0.254	0.354
COMET-QE-DA	0.29	0.34	0.08	0.07	0.3	0.45	0.27	0.33	0.235	0.298
COMET-QE-MQM	0.23	0.35	0.13	0.2	0.3	0.44	0.22	0.29	0.22	0.32
COMET-Kiwi	0.344	0.475	0.115	0.129	0.371	0.514	0.322	0.392	0.288	0.378
COMET-Kiwi-xl	0.334	0.48	0.3	0.338	0.337	0.486	0.266	0.352	0.309	0.414
GEMBA-MQM	0.235	0.266	0.085	0.118	0.108	0.079	0.282	0.235	0.178	0.174
GEMBA-MQM(IL lang)	0.228	0.276	0.081	0.077	0.05	0.069	0.171	0.261	0.132	0.171

What about IndicCOMET models in zero-shot style?

Metric	asm		mai		kan		pan		Average	
	τ	ρ	τ	ρ	τ	ρ	τ	ρ	τ	ρ
Indic-COMET-DA	0.263	0.348	0.221	0.3	0.353	0.511	0.293	0.361	0.283	0.38
Indic-COMET-MQM	0.201	0.270	0.201	0.288	0.251	0.388	0.282	0.340	0.234	0.322
Base-IndicBERT(DA)	0.273	0.396	0.380	0.552	0.384	0.528	0.259	0.353	0.324	0.457
Base-IndicBERT(MQM)	0.293	0.426	0.311	0.483	0.302	0.44	0.224	0.313	0.283	0.416

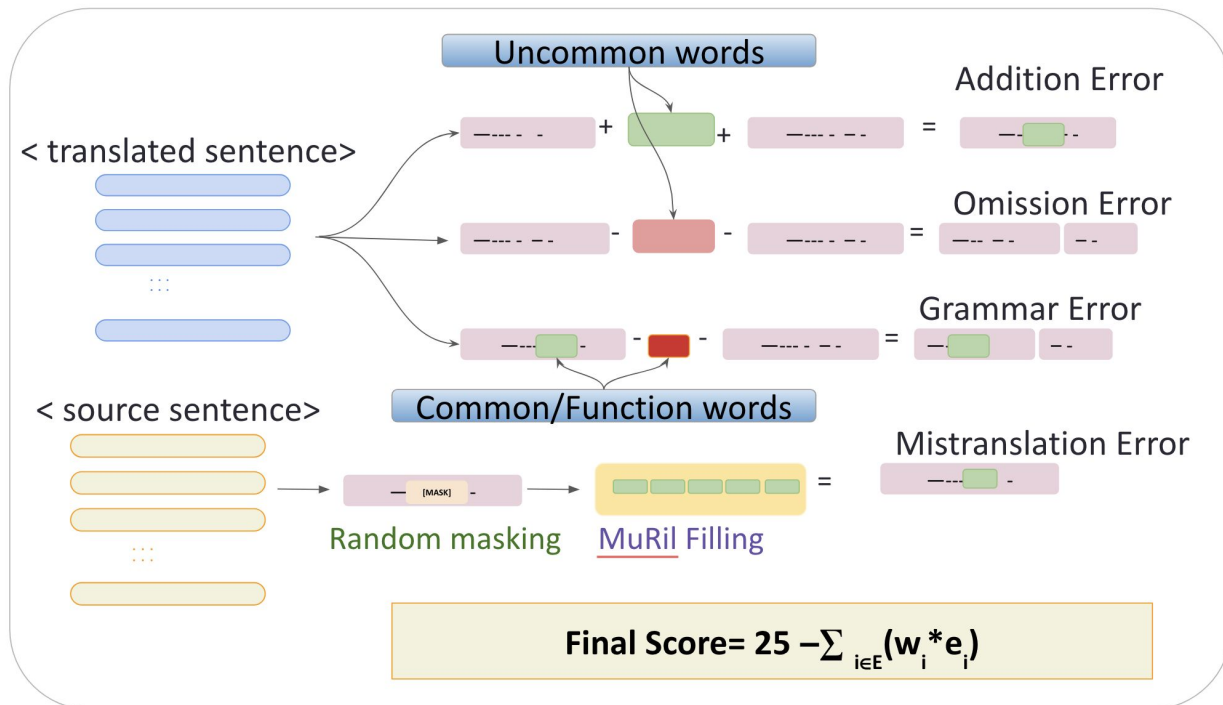
- COMET-DA and MQM models as well as IndicBERT models fine-tuned
- Best average Kendall Tau and Pearson scores of 0.537 and 0.393 for supervised
- Drops to 0.324 and 0.457 for new languages zero-shot
- Lesson: Careful how you use learned metrics on unseen languages

No Data? No Problem!

Just Make Synthetic Data!

Methodology

- BPCC seed parallel dataset to generate synthetic data
- Error types and severities based on distributions from the related languages of supervised dataset.
- Created synthetic data: 44k Assamese, 32k Kannada, 24k Maithili, and 6k Punjabi sentences.
- Introduced errors such as addition, omission, mistranslation, and grammar in the original translations.



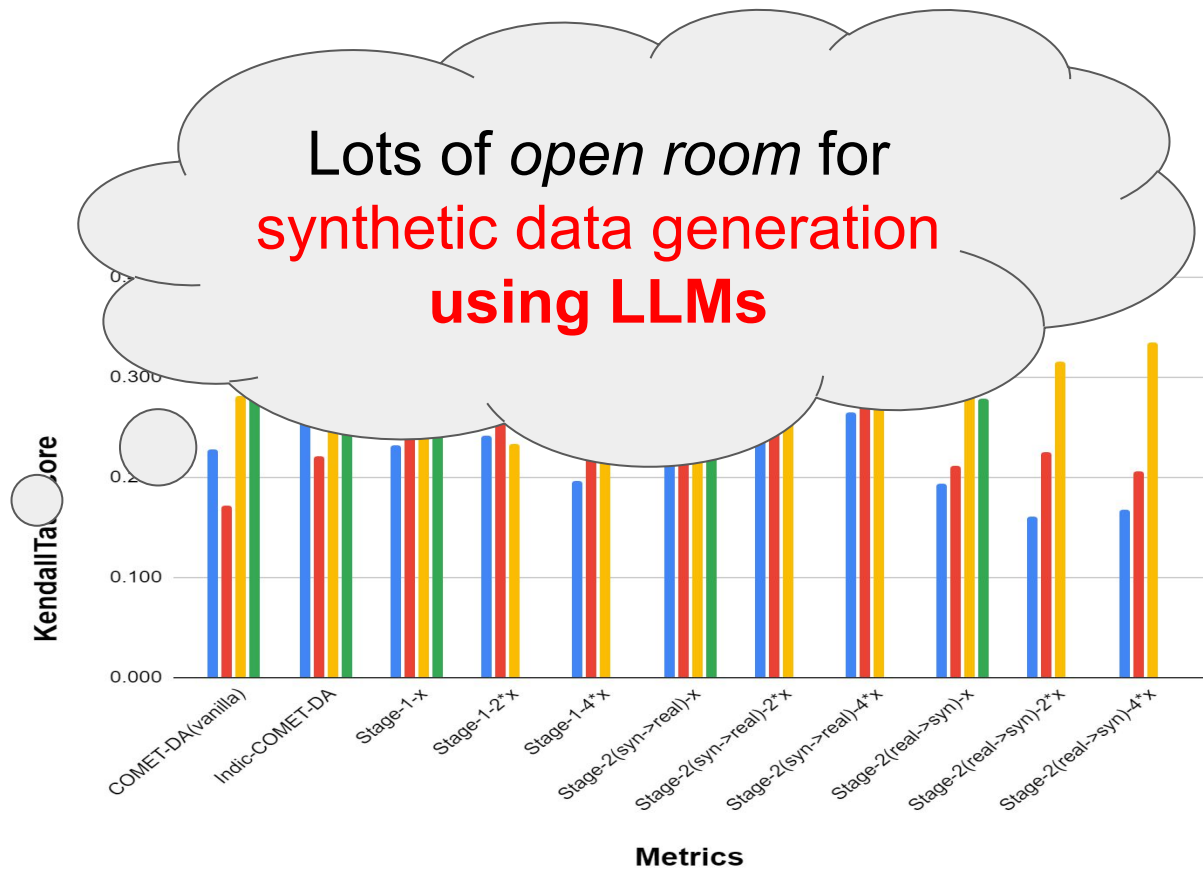
Settings

- **Joint training** of synthetic data and supervised data of related languages
 - Related languages' data = 5000 instances for 5 languages ($X=5000$)
 - Synthetic data of various scales: X , $2X$, $4X$
 - Synthetic data of only the relevant language

- **2 Stage training:**
 - Synthetic \rightarrow Related
 - X , $2X$, $4X \rightarrow X$
 - Related \rightarrow Synthetic
 - $X \rightarrow X$, $2X$, $4X$

Does Synthetic Data Help?

- Not quite!
- Only Maithili benefits from synthetic data
 - Currently investigating why
- Increasing synthetic data doesn't necessarily benefit
- Joint vs stagewise didnt give any satisfactory conclusion



Summary

- IndicBART and Low-Resource MT
 - Showing that language family specific models are crucial
- IndicTrans2
 - Supporting all 22 scheduled languages with language family focused efforts
- Indic MT Using LLMs
 - Neural re-rankers and leverage monotonicity, leveraging transliteration
- IndicMT Eval
 - Choose your metrics wisely and make IndicComet

Future Work

- Scale up monolingual data and identify how to obtain good synthetic data
 - How to overcome translationese effects?
- Larger models: IndicBART is due for an upgrade
 - Scaling to larger number of parameters
- Improved neural rerankers for retrieval augmented MT and Evaluation
 - Language aware retrieval
- Better evaluation metrics, synthetic data creation
 - LLMs as error span annotators
- Support for dialects and endangered languages
 - Involve your family :)
- Speech translation
 - Rather underexplored for Indic languages

Special Thanks to AI4Bharat,
IITB Collaborators

Q/A