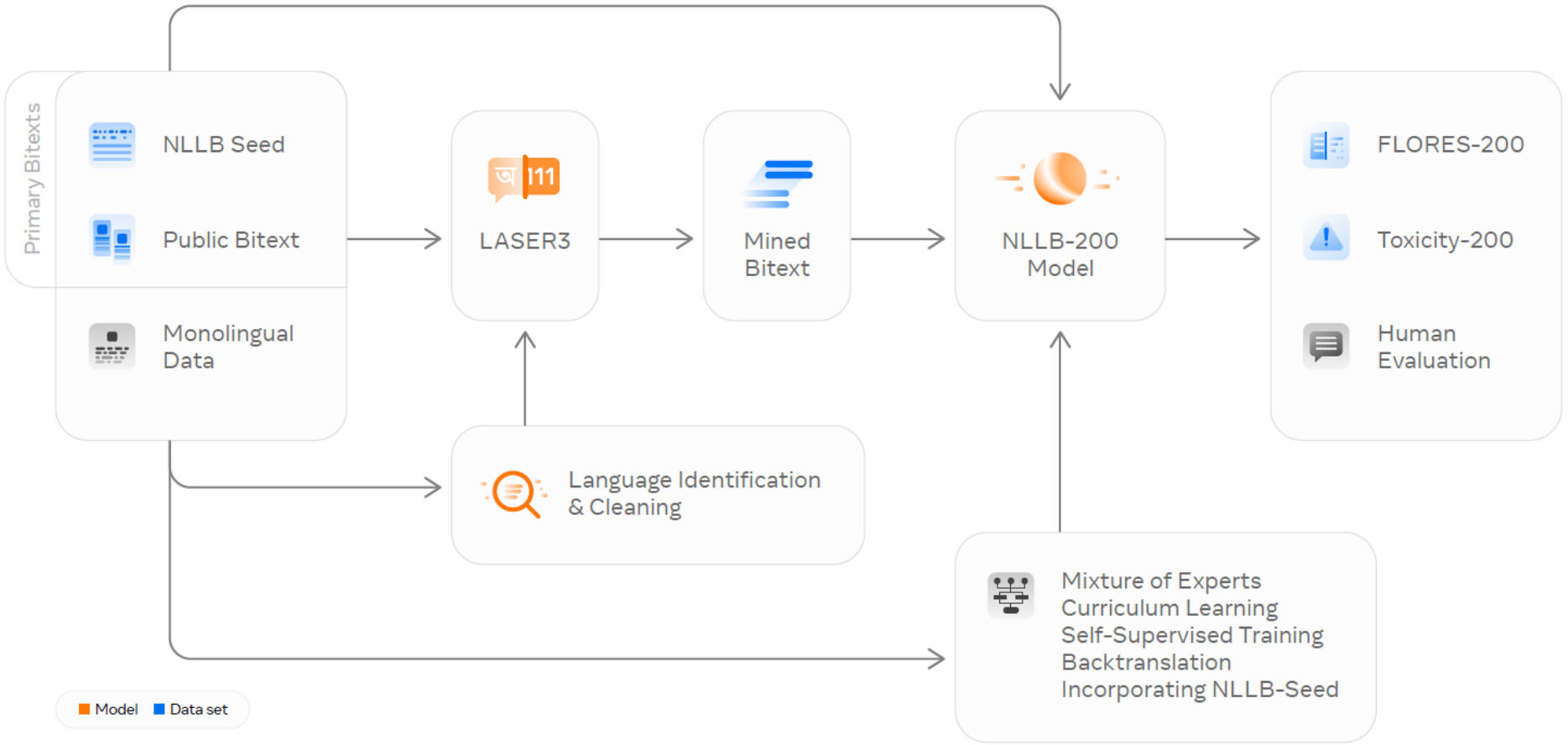
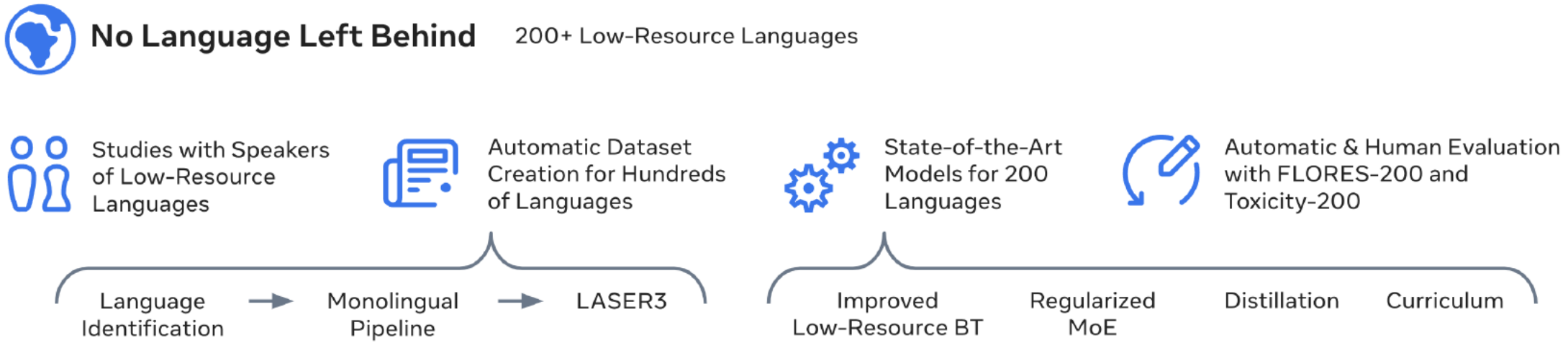
Machine Translation

No Language Left Behind

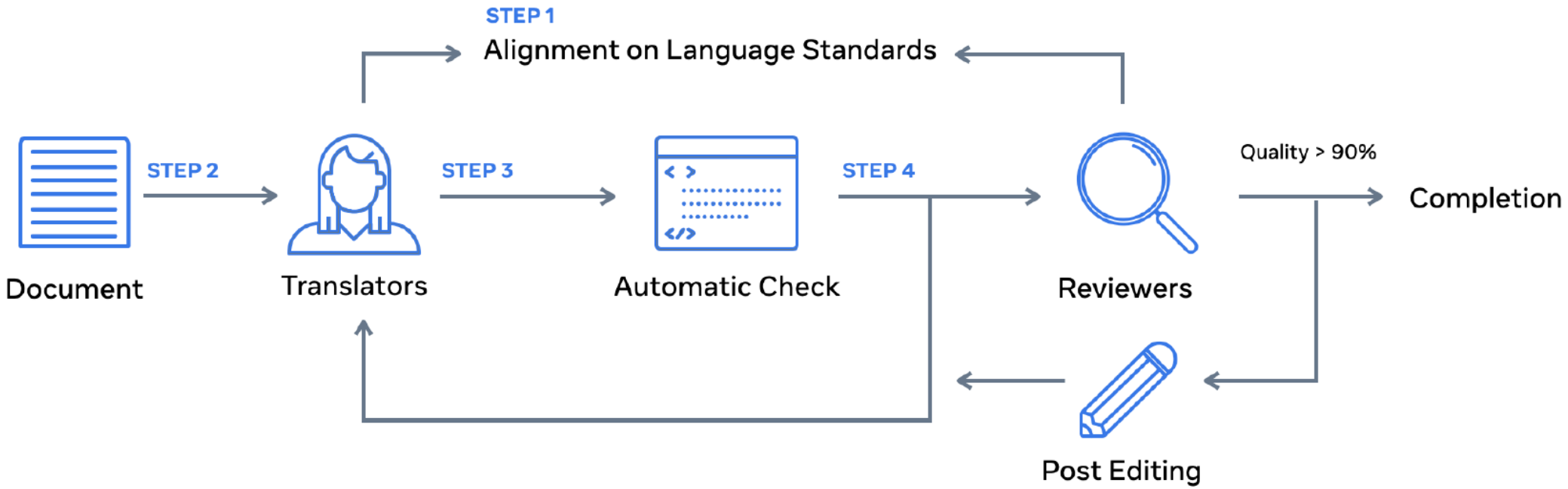


No Language Left Behind



Translation Data

Manual Data Curation



Manual Data Curation - Statistics

Overview Statistics	
# of sentences	3001
Avg # of words/sentence	21
# of articles	842

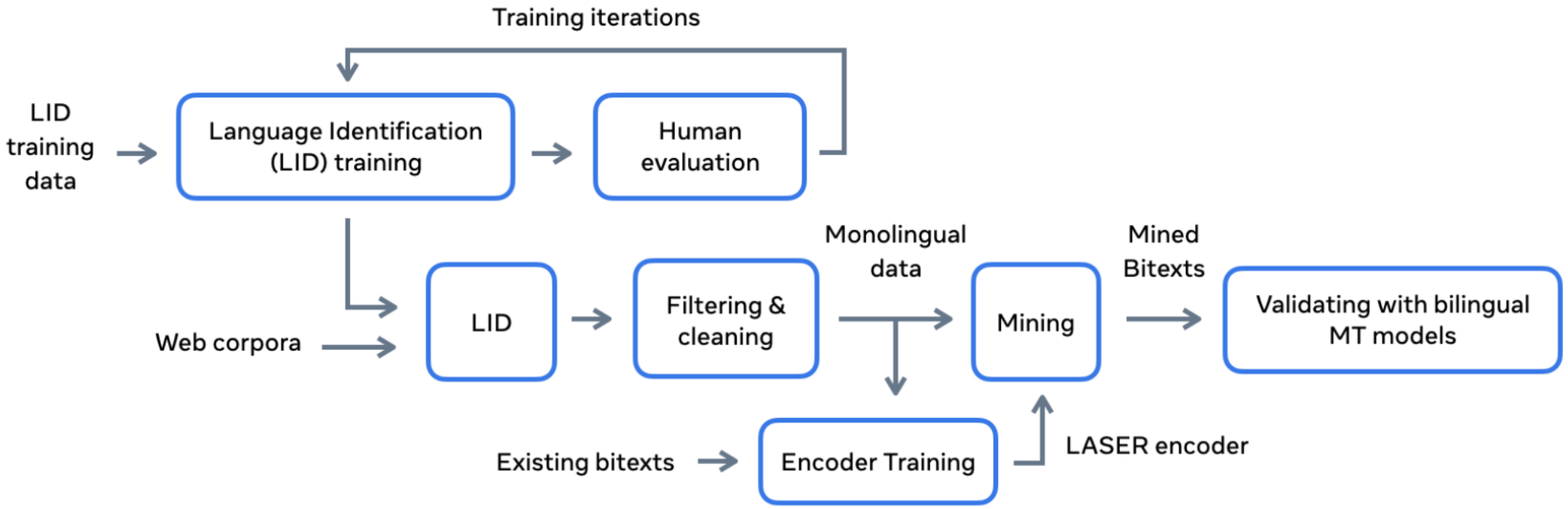
Split	**# of sentences**
dev	997
devtest	1012
test	992

# of Languages requiring Re-translation	10
Avg # of Re-translations	1
Max # of Re-translations	2

Avg # of Days to Translate 1 language	42
Avg # of Days to align	28
Avg # of Days for 1 language	119
Shortest Turnaround (days) for 1 language	70
Longest Turnaround (days) for 1 language	287

Automated Data Curation

Broad Pipeline



Language Identification (LID)

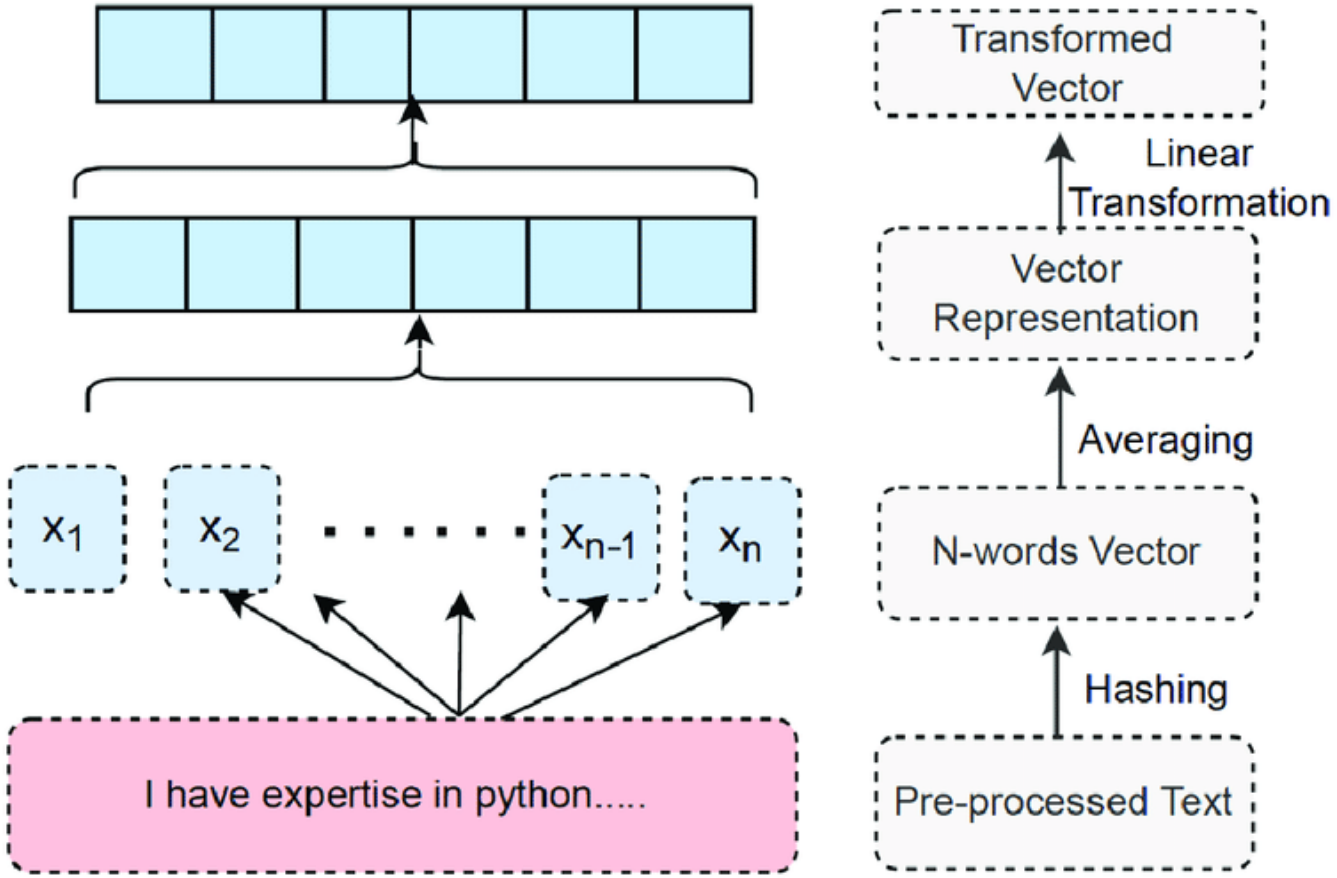
- Models to detect the language of the input text

- Considerations:
  - Training data needs to be clean and cover multiple domain
  - Need to be light-weight to handle large scale data

Data Cleaning For LID

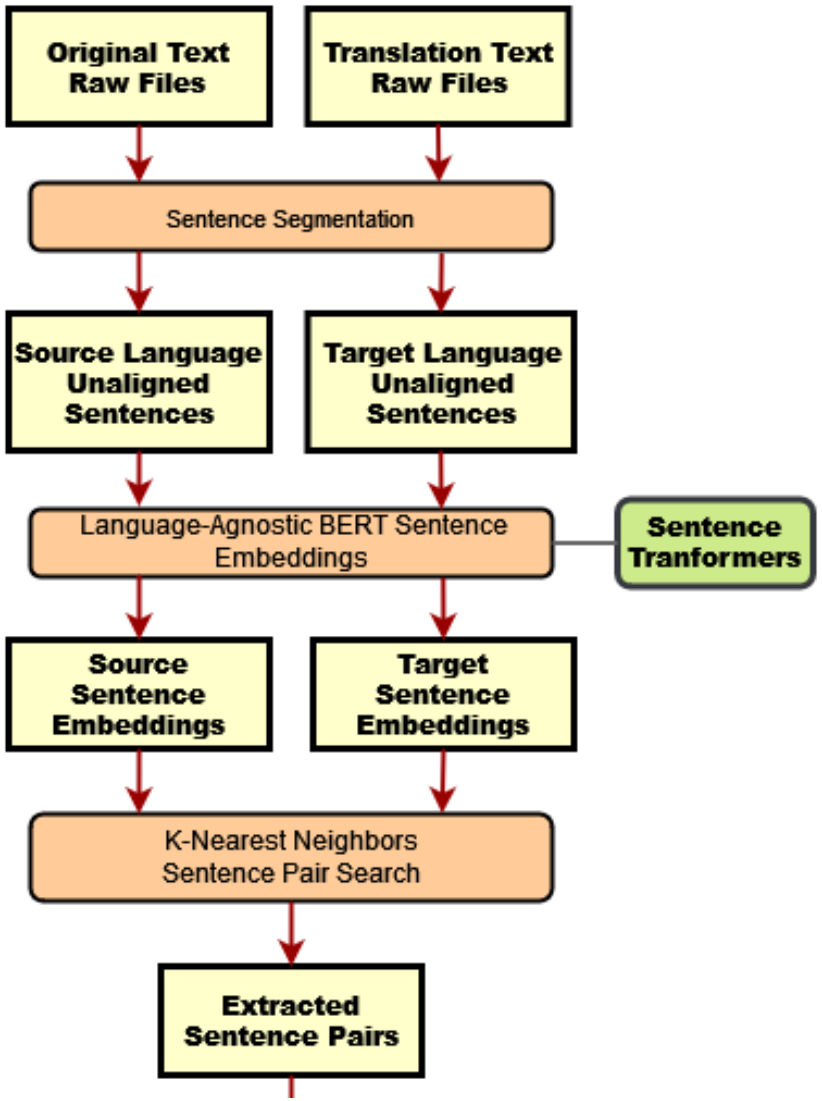
Filter	Label	Filtered Sentence
Histogram	urd_Arab	H█═══𝓜😊♓𝓜 ɑ Đ ɑ█ مرتضےپاشاےیا :
	dan_Latn	అనంతపూర్ డిస్ట్రిక్ , urdu: آدل آباد ضلع) er et distrikt i den
Script	jpn_Jpan	4.0, CUDA 対応。消費電力は 40W。Quadro FX 380 コア 450MHz
	zho_Hant	容存档于 2009 年 2 月 10 日）． Satellite map 維基衛星
English	tur_Latn	A module is said to be semisimple if it is the sum of simple submodules.
	nld_Latn	Line drawing and design: From the book Brazil and the Brazilians, 1857

Fasttext Model For LID

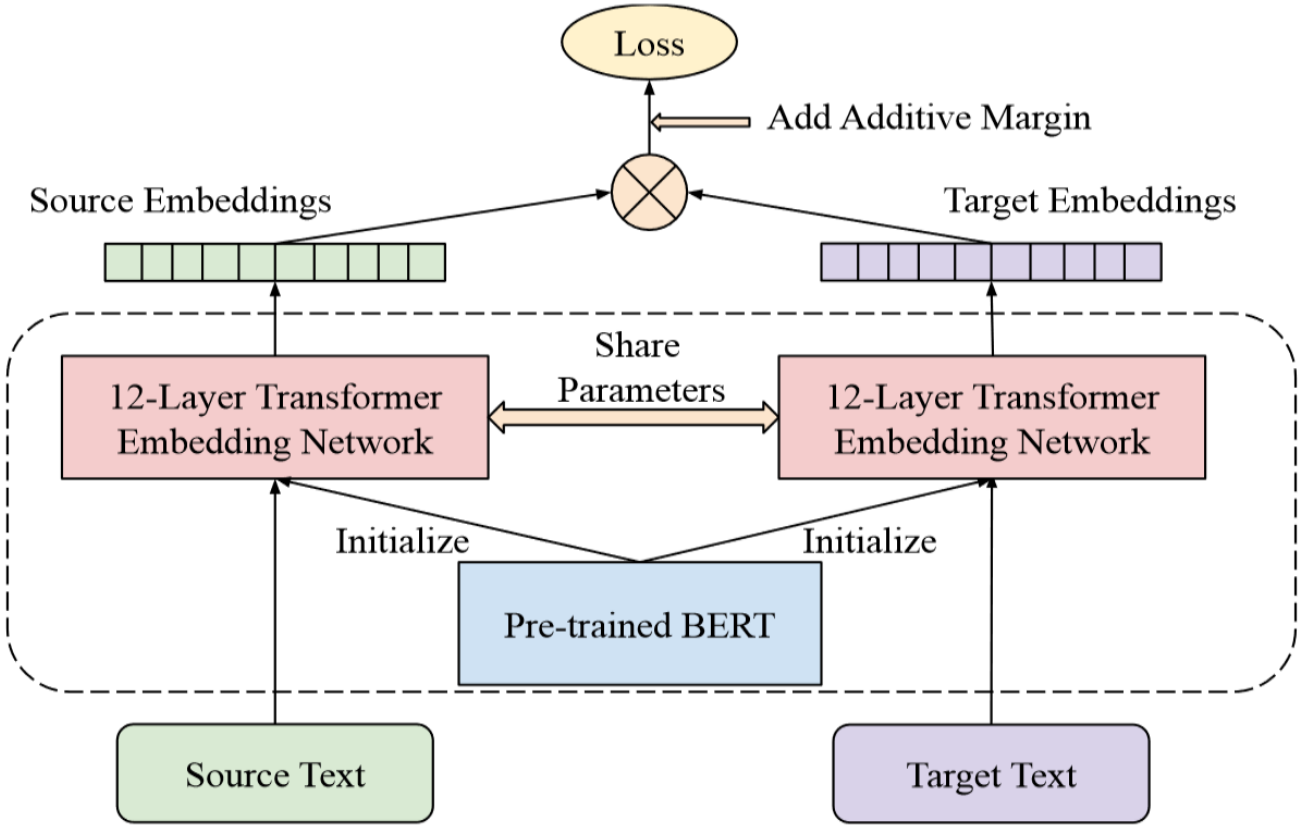


Bitext Mining

- Language agnostic representations are of raw sentences obtained
- Cosine similarity is utilized to obtain a measure of cross-lingual similarity
- Typically utilize scalable and efficient vector search engines like FAISS



Language Agnostic Encoders



Data Filtering For Encoders

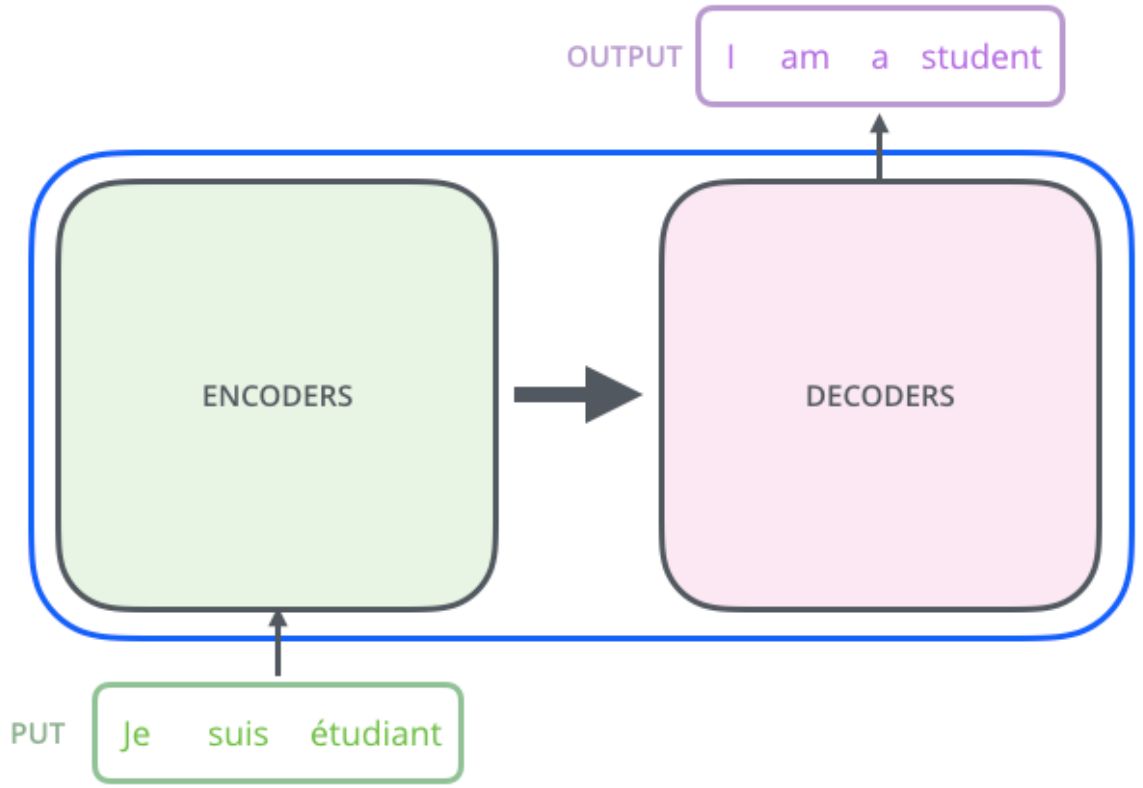
Filter	Example	Reason
Low LID threshold	Internet Plus € 58,50	`eng_Latn` at 0.19 LID score
LID mismatch	Best véto ever!	doc. LID French, sent. LID Czech
Numbers	Vol.180 Sep. (2011)	exceeded numbers ratio
Punctuation	. * sApEvAte cHe... » (Previous page)	exceeded punctuation ratio
Emoji	💪💪💪 #gymgirl	exceeded emoji ratio

Translation Modelling

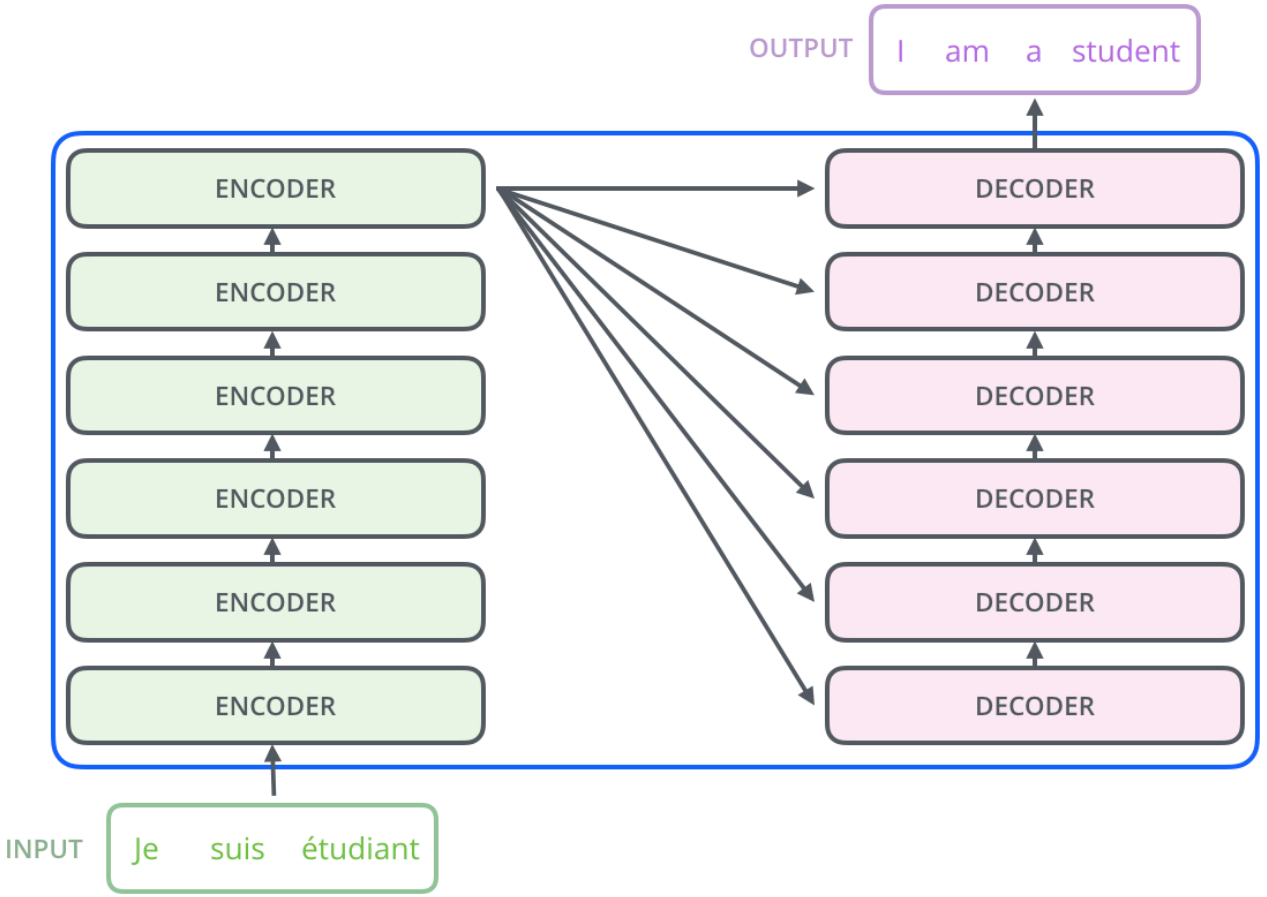
Transformer Architecture

- Transformer architecture was introduced for seq2seq tasks, specifically machine translation

- Encoders takes as input source sentence and decoder auto-regressively generates the target sentence



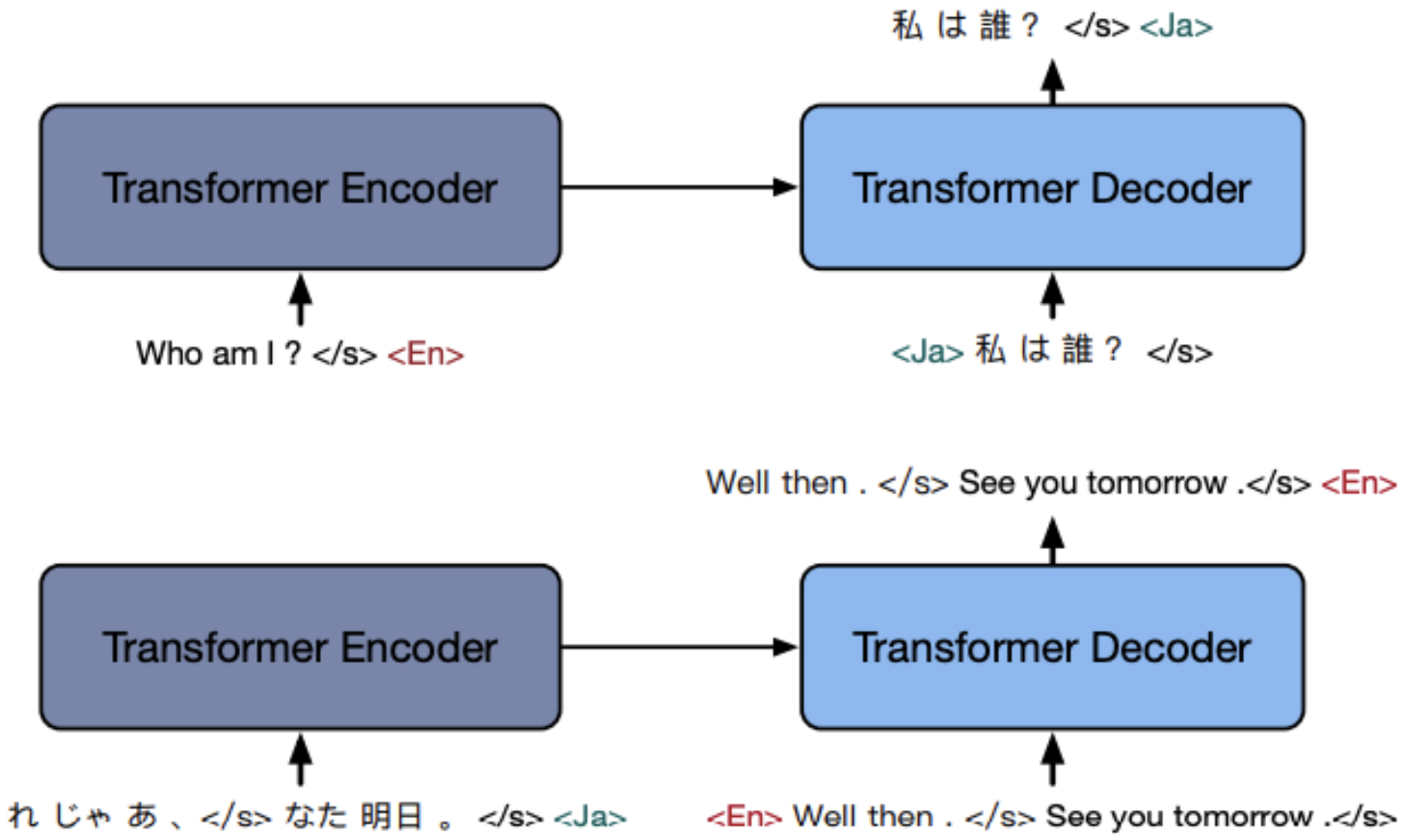
Scaling Encoder - Decoder



Under The Hood: Decoder

- Sample Input: A B C D E F
- Model inputs per iteration:
  - A B
  - A B C Attention mask: 1 1 1 0 0 0
  - A B C D Attention mask: 1 1 1 1 0 0
  - A B C D E
  - A B C D E F

Multilingual Transformers



Model Considerations

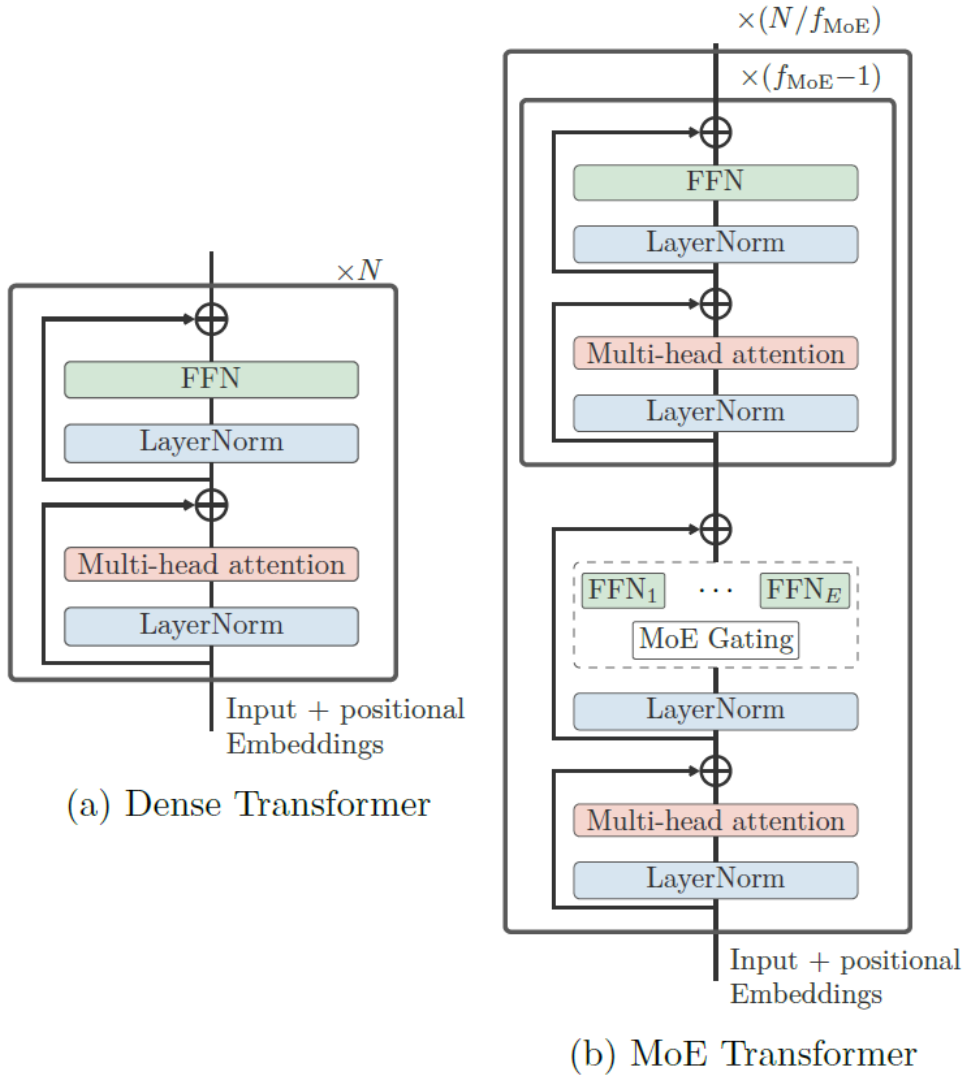
- **Vocabulary Size:** A vocabulary size increase from 32k to 64k did not provide noticeable score improvement but drastically slowed generation

- **Encoder-Decoder Size:** A deep-decoder causes increase in training and inference time, a deep-encoder and shallow-decoder generates high quality translation and faster training and inference

- **Multilingual Models:** Modelling several similar languages together leads to improved performance across all these languages

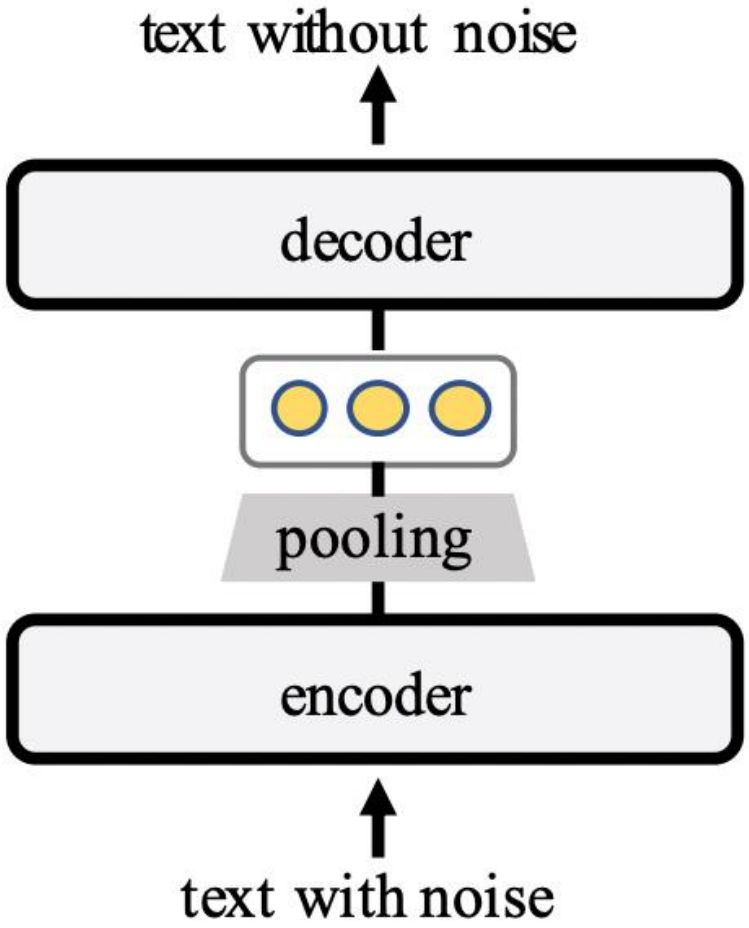
Mixture Of Experts

- The capacity of a neural network to absorb information is limited by the number of its parameters
- More parameters equals better capacity
- MOE is a type of conditional computation where parts of the network are activated on a per-example basis



Pretraining Tasks

- **Denoising Auto Encoder**
  - Predicting a corrected text from a corrupted input text
- **Causal Language Modelling**
  - Language modelling task where source is empty, and target is text from monolingual corpus



Curriculum Learning

- Pretraining on SSL, followed by finetuning on MMT (DAE -> MMT)
- Multitask training on SSL and MMT (DAE + MMT)
- Multitask training on SSL and MMT, followed by finetuning on MMT (DAE + MMT -> MMT)

Curriculum Learning Results

	eng_Latn-xx				xx-eng_Latn				xx-yy
	all	high	low	v.low	all	high	low	v.low	all
MMT	43.3	**55.4**	38.4	31.6	53.5	**63.6**	49.4	46.5	41.3
DAE⇒MMT	42.6	55.0	37.6	30.8	52.3	62.2	48.3	45.4	40.4
DAE+MMT	**43.5**	55.2	**38.8**	**32.7**	**54.4**	**63.6**	**50.7**	**48.4**	**42.4**
DAE+MMT⇒MMT	43.4	**55.4**	38.5	32.2	54.3	**63.6**	50.5	48.0	42.2

Pretraining Tasks Impact

	eng_Latn-xx				xx-eng_Latn				xx-yy
	all	high	low	v.low	all	high	low	v.low	all
MMT	43.3	**55.4**	38.4	31.6	53.5	**63.6**	49.4	46.5	41.3
MMT+LM	42.6	54.9	37.5	30.8	53.5	**63.6**	49.4	46.7	41.5
MMT+DAE	**43.5**	55.2	**38.8**	**32.7**	**54.4**	**63.6**	**50.7**	**48.4**	**42.4**
MMT+DAE+LM	42.6	55.0	37.6	31.4	53.4	62.7	49.6	47.0	40.8

Data Augmentation

Sources Of Data

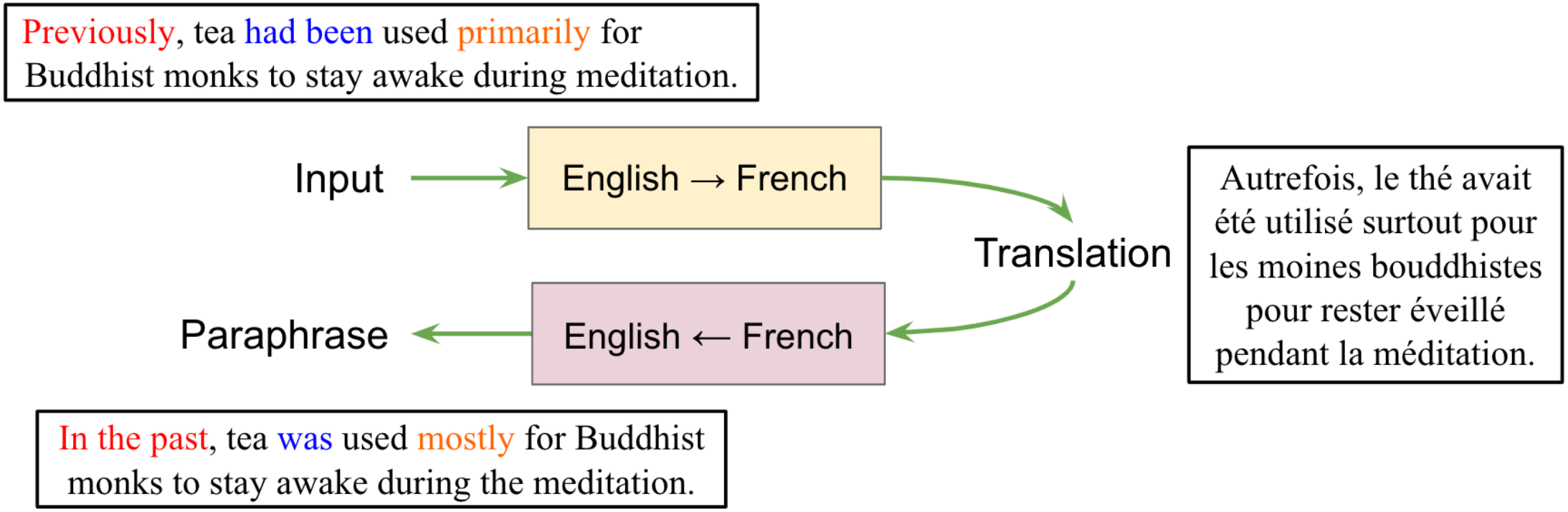
Source	Human Aligned?	Noisy?	Limited Size?	Model-Dependent?	Models Used
NLLB-SEED	✓	✗	✓	✗	—
PUBLICBITEXT	✗	✓	✓	✗	—
MINED	✗	✓	✗	✓	Sentence Encoders
MMTBT	✗	✓	✗	✓	Multilingual
SMTBT	✗	✓	✗	✓	Bilingual MOSES
Ideal Data	✓	✗	✗	✗	—

Impact Of Data Sources

	eng_Latn-xx				xx-eng_Latn				xx-yy
	all	high	low	v.low	all	high	low	v.low	all
PRIMARY	41.0	52.8	36.3	28.1	47.4	60.5	42.1	36.7	39.2
+MINED	43.8	55.2	39.2	**34.0**	53.9	64.4	49.6	46.1	40.9
+MMTBT	44.0	55.1	39.5	**34.0**	55.7	64.8	52.0	50.8	40.6
+SMTBT	**44.2**	**55.5**	**39.6**	**34.0**	**55.9**	**64.9**	**52.2**	**50.9**	**41.1**

- Note: Use data tag such as <MMT_BT_DATA> or <MINED_DATA> to help model discern the data sources

Backtranslation



Knowledge Distillation

Distilling Knowledge From Larger Model

		eng_Latn-xx				xx-eng_Latn				xx-yy	Avg.
	size	all	high	low	v.low	all	high	low	v.low	all	all
NLLB-200	54B	45.3	54.9	41.9	39.5	56.8	63.5	54.4	54.4	42.7	48.3
dense baseline	1.3B	43.5	52.8	40.1	37.6	54.7	61.8	52.2	51.9	41.0	46.4
dense distilled	1.3B	**44.0**	**53.2**	**40.8**	**38.4**	**55.1**	**61.9**	**52.6**	**52.5**	**41.5**	**46.9**
dense baseline	615M	41.4	50.7	38.1	35.1	52.2	59.7	49.6	49.1	39.3	44.3
dense distilled	615M	**41.8**	**50.9**	**38.5**	**35.8**	**52.3**	59.7	**49.7**	**49.3**	**39.5**	**44.6**