# Hugely Multi-Lingual Models

Department: name: Speech and Semantic Lab, Noah’s Ark Lab
Author’s name: Valentin Malykh
Date: 20/09/2022

HUAWEI

# Machine Translation


HUAWEI

# Agenda

- Papers
- Data
- Models
- Evaluation

HUAWEI

# Papers

- **No Language Left Behind: Scaling Human-Centered Machine Translation.**
  - 06.2022
  - 190 pages
  - 38 authors
  - 14 first authors and 14 last authors
  - 37 from Meta AI Research and 1 from UC Berkeley
  - *https://research.facebook.com/file/585831413174038/No-Language-Left-Behind--Scaling-Human-Centered-Machine-Translation.pdf*

- Building Machine Translation Systems for the Next Thousand Languages.
  - 05.2022
  - 77 pages
  - 24 authors
  - 2 first authors
  - All from Google
  - *https://arxiv.org/pdf/2205.12654*

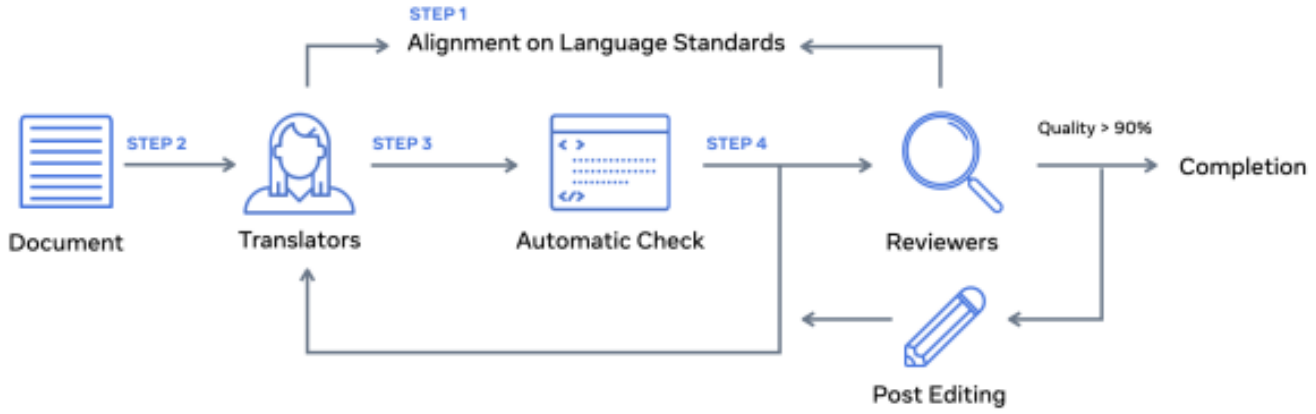
HUAWEI

# No Language Left Behind: Scaling Human-Centered Machine Translation

# Data / Human Translated

- Flores-101
  - 101 languages
  - 3001 sentences from English Wikipedia

- Flores-200
  - 204 languages
  - The same 3001 sentences from English Wikipedia




HUAWEI

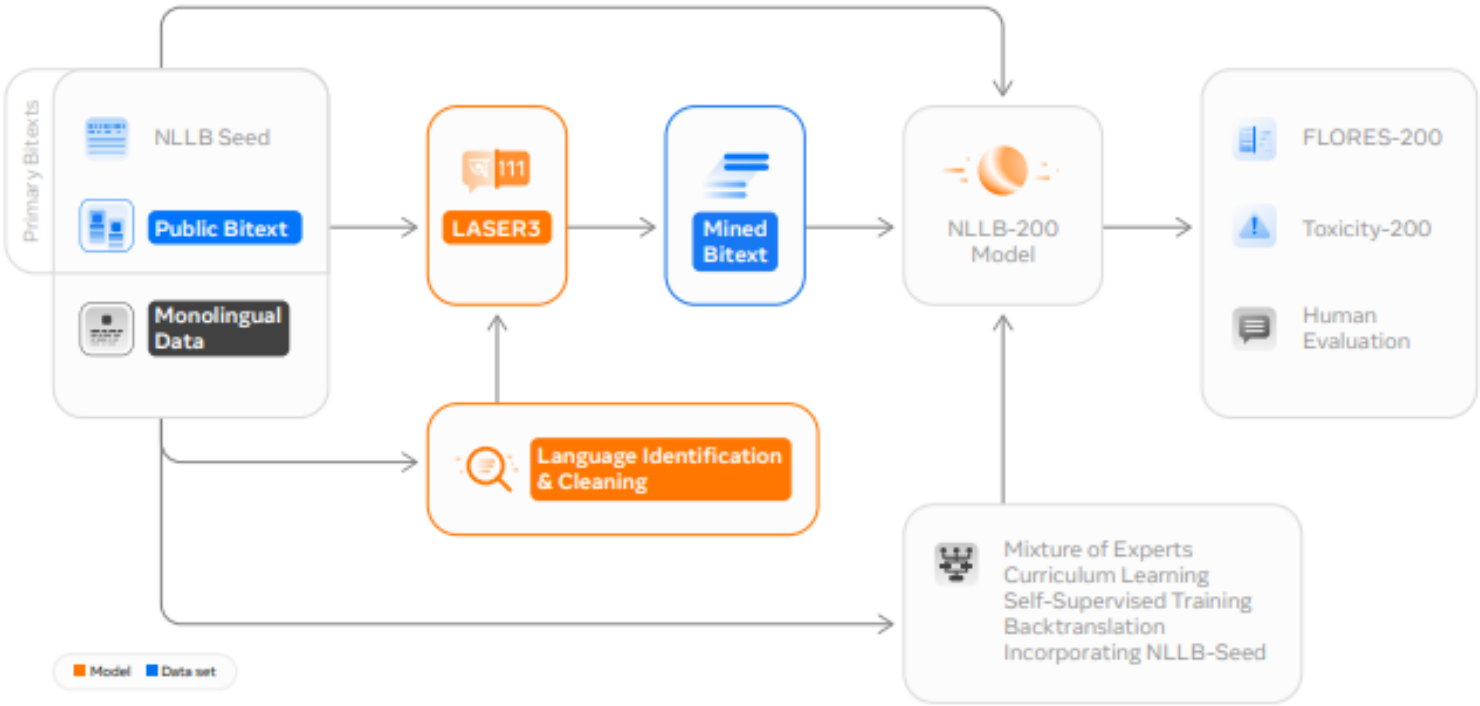
# Data / Human Translated (2)

- NLLB-Seed
  - 39 languages
  - 6000 sentences from good Wikipedia Articles
  - No human verification

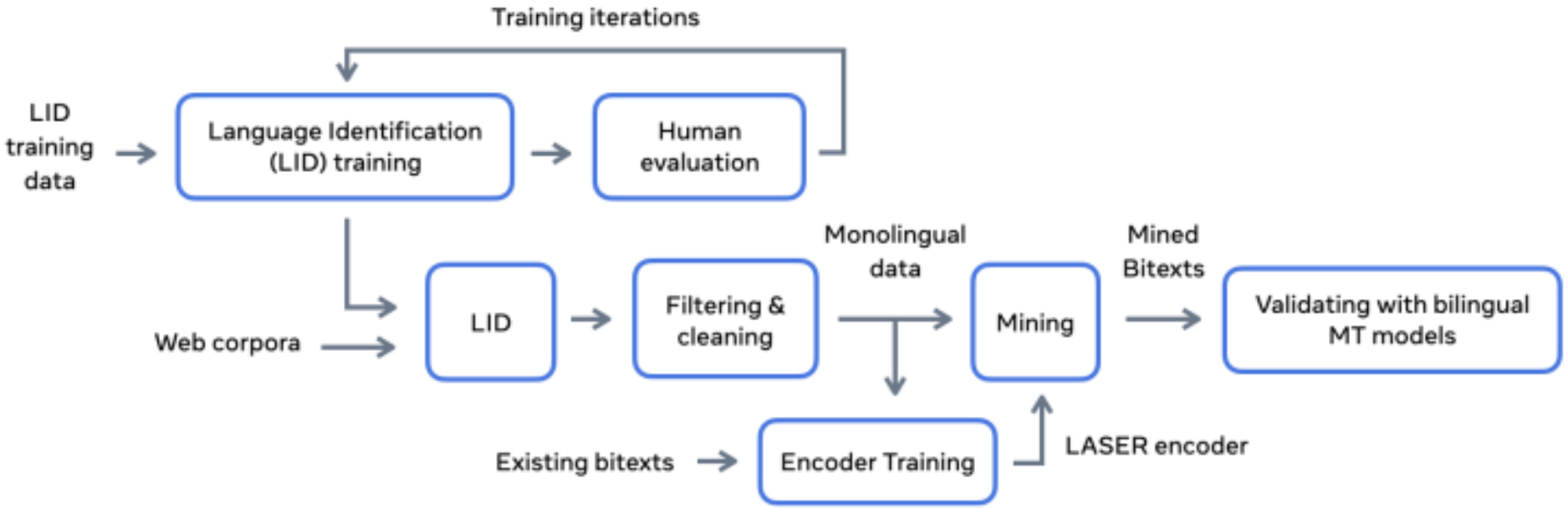
- NLLB-MD
  - 6 languages
  - 4 domain (News, Health, Formal Speech, Informal Speech)
  - 3000 sentences

HUAWEI

# Data / Mined


Primary Bitexts
NLLB Seed
Public Bitext
Monolingual Data
LASER3
Mined Bitext
NLLB-200 Model
FLORES-200
Toxicity-200
Human Evaluation
Language Identification & Cleaning
Mixture of Experts
Curriculum Learning
Self-Supervised Training
Backtranslation
Incorporating NLLB-Seed
Model
Data set

# Data / Mined (2)



# Data / Mined (3)

- fasttext
- Trained on NLLB-Seed

| | # Supported Languages | FLORES-200 ∩ CLD3 ∩ LangId ∩ LangDetect 51 Labels | | FLORES-200 ∩ CLD3 ∩ LangId 78 Labels | | FLORES-200 ∩ CLD3 95 Labels | |
|---|---|---|---|---|---|---|---|
| | | **F1** | **FPR** | **F1** | **FPR** | **F1** | **FPR** |
| LangDetect | 55 | 97.3 | 0.0526 | 64.4 | 0.4503 | 53.1 | 0.4881 |
| LangId | 97 | 98.6 | 0.0200 | 92.0 | 0.0874 | 75.8 | 0.2196 |
| CLD3 | 107 | 98.2 | 0.0225 | 97.7 | 0.0238 | 97.0 | 0.0283 |
| Ours | **218** | **99.4** | **0.0084** | **98.8** | **0.0133** | **98.5** | **0.0134** |

Table 4: **Comparison of Open-Source Language Identification Models with various intersections of labels.** F1 is the micro F1 score and FPR is the micro False Positive Rate.

| | **Micro F1** | **Macro F1** | **Macro Precision** | **Macro Recall** | **Macro FPR** | **Micro FPR** |
|---|---|---|---|---|---|---|
| Low-Resource | 95.63 | 95.9 | 97.6 | 95.4 | 0.01213 | 0.0235 |
| All FLORES-200 | 95.85 | 95.5 | 94.0 | 95.7 | 0.02110 | 0.0210 |

Table 5: **Performance of our LID system on FLORES-200**. Arabic languoids and Akan/Twi have been merged after linguistic analysis.

# Data / Mined (4)

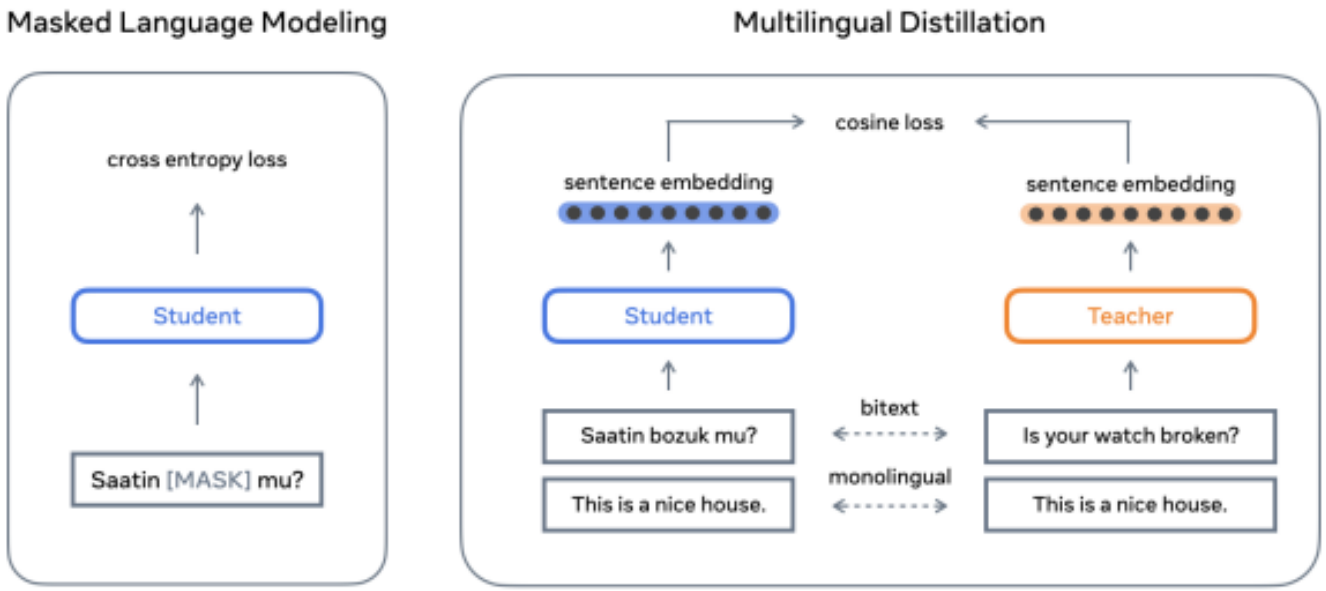
- Common Crawl and ParaCrawl
  - Like in earlier CCNet and CCMatrix
- Language Identification filtering
  - If a sentence language does not match a document language, discard
- Heuristic filtering
  - >20% punctuation, URLs, etc.
- Deduplication
- Language Model filtering
  - Is applied only to English data, but due to bitext mining it is ok

| | Total | Min | | Max | | Median | | Mean |
|---|---|---|---|---|---|---|---|---|
| **Low-Resource** | | | | | | | | |
| Raw Data (Para.) | 2.4B | 27.1K | tzm_Tfng | 465.8M | nob_Latn | 3.3M | mai_Deva | 17.9M |
| LID/Script Mismatch (Sent.) | 0.3B | 0.2K | tzm_Tfng | 47.8M | nob_Latn | 0.3M | fao_Latn | 2.2M |
| Clean Sentences | 3.6B | 1.3K | tuk_Latn | 330.3M | glg_Latn | 4.4M | tso_Latn | 26.8M |
| **High-Resource** | | | | | | | | |
| Raw Data (Para.) | 105.4B | 4247.8K | tsn_Latn | 73.2B | eng_Latn | 83.1M | eus_Latn | 3401.2M |
| LID/Script Mismatch (Sent.) | 2.5B | 124.4K | ben_Beng | 1.6B | eng_Latn | 5.0M | als_Latn | 81.2M |
| Clean Sentences | 40.1B | 5153.8K | tsn_Latn | 21.5B | eng_Latn | 234.3M | als_Latn | 1294.5M |

HUAWEI

# Data / LASER3

- LASER model (teacher)
  - 93 languages from OPUS
  - 5 layer biLSTM + maxpooling
  - Sentencepiece
- LASER3 model (student)
  - 12 layer transformer
  - 1024 hidden
  - 250M params

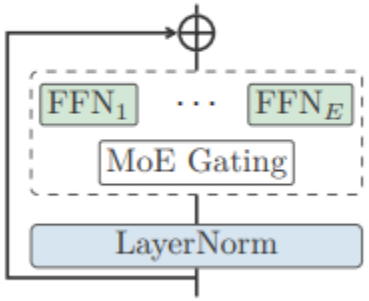


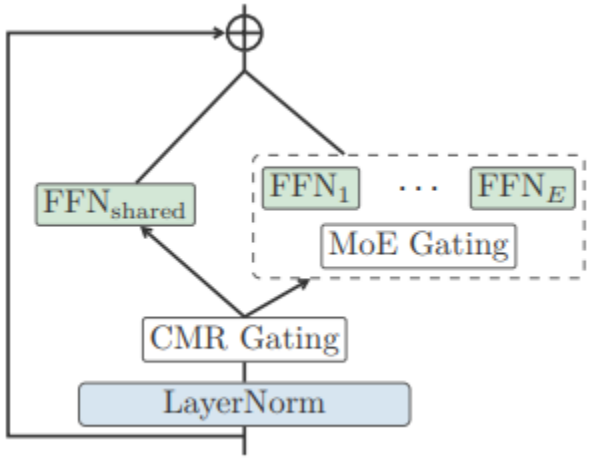

$$\text{score}(x, y) = \text{margin}\left(cos(x, y), \sum_{z \in NN_k(x)} \frac{cos(x, z)}{2k} + \sum_{v \in NN_k(y)} \frac{cos(y, v)}{2k}\right)$$

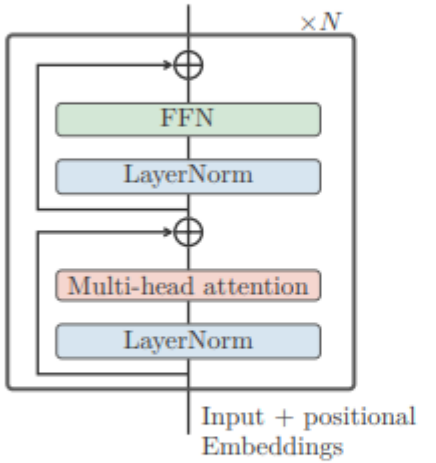
HUAWEI

# Model / Details

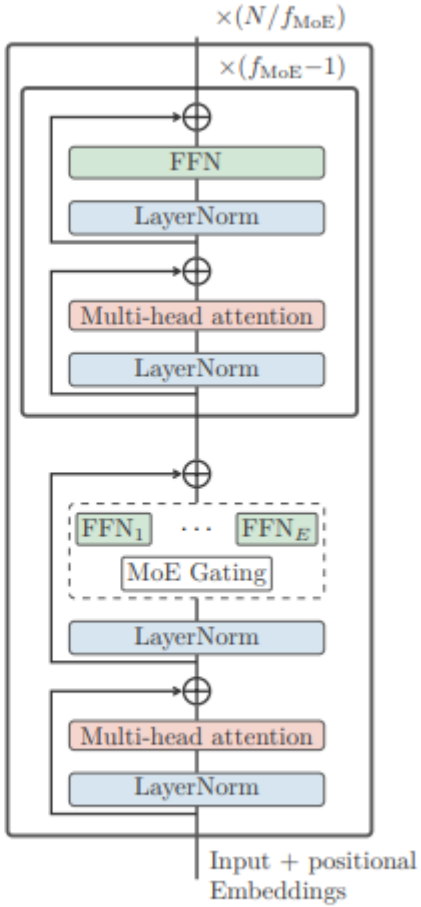
- **NLLB-200**
  - Mixture of Experts
  - Curriculum learning
    - Denoising Autoencoder (+)
    - Causal Language Modeling (-)
  - Self-Supervised Learning
  - Back-Translation




(a) Dense Transformer




(b) MoE Transformer

| | eng_Latn-xx | | | | xx-eng_Latn | | | | xx-yy |
|---|---|---|---|---|---|---|---|---|---|
| | all | high | low | v.low | all | high | low | v.low | all |
| MoE-64 $p_{\text{drop}}$=0 | 43.3 | 55.9 | 38.2 | 29.6 | 52.9 | 63.9 | 48.4 | 43.7 | 39.3 |
| MoE-64 $p_{\text{drop}}$=0.3 † | 44.3 | 56.0 | 39.5 | 32.5 | 54.4 | 63.9 | 50.6 | 47.7 | 41.9 |
| MoE-64 FOM ($p_{\text{drop}}$=0.3, $p_{\text{fom}}$=0.1) † | 43.8 | 55.6 | 38.9 | 32.5 | 54.8 | 64.3 | 50.9 | **48.5** | 42.0 |
| MoE-64 EOM ($p_{\text{drop}}$=0.3, $p_{\text{eom}}$=0.1) † | **44.7** | **55.9** | **40.1** | **33.4** | 54.8 | 64.3 | 51.0 | 48.3 | **42.5** |
| MoE-64 CMR ($p_{\text{drop}}$=0.2, $p_{\text{cmr}}$=0.2) † | **44.8** | **56.1** | **40.2** | **33.4** | **55.2** | **64.6** | **51.4** | **48.5** | **42.6** |
| Gating Dropout($p_{\text{drop}}$=0.3, $p_{gd}$=0.2) † | 44.4 | 55.7 | 39.8 | 33.0 | 54.8 | 64.1 | 51.0 | **48.5** | 42.3 |

# Model / Details (2)

- NLLB-200
  - Transformer encoder-decoder
  - replace FFN in every 4-th Transformer block with a Sparsely Gated Mixture of Experts layer containing 128 experts
  - model dimension 2048
  - FFN dimension 8192
  - 16 attention heads
  - 24 encoder layers
  - 24 decoder layers
  - Shared embedding weights
  - EOM with peom=0.2.

  - The model has a total of **54.5B** parameters and FLOPs similar to that of a **3.3B** dense model.

HUAWEI

# Model / Dataset

- Primary Bitext
  - NLLB-Seed
  - 661 en-xx MT dataset
- Mined Bitext
  - 784 directions
  - xsim < 5
- Monolingual data
  - 192 languages

HUAWEI

# Evaluation / Metrics

- BLEU
- spBLEU
  - Based on sentencepiece
- chrF++
  - Symbol n-grams & token unigrams and bigrams

HUAWEI

# Evaluation / Results

- Comparison with enterprise systems
- Comparison
  - **spBLEU** / chrF++

| | eng_Latn-xx | | | | | xx-eng_Latn | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | (a) | (b) | (c) | (d) | NLLB-200 | (a) | (b) | (c) | (d) | NLLB-200 |
| asm_Beng | 6.9/- | -/- | -/- | 24.9/- | **27.6**/35.9 | 23.3/- | -/- | -/- | 13.6/- | **36.1**/57.8 |
| ben_Beng | 20.3/- | 17.3/- | 34.6/- | 31.2/- | **37.9**/50.0 | 32.2/- | 30.7/- | 28.1/- | 22.9/- | **41.2**/62.2 |
| guj_Gujr | 22.6/- | 22.6/- | 40.2/- | 35.4/- | **41.4**/53.3 | 34.3/- | 33.6/- | 25.6/- | 27.7/- | **46.8**/66.6 |
| hin_Deva | 34.5/- | 31.3/- | **44.2**/- | 36.9/- | 42.6/57.3 | 37.9/- | 36.0/- | 38.7/- | 31.8/- | **46.5**/66.5 |
| kan_Knda | 18.9/- | 16.7/- | 32.2/- | 30.5/- | **40.8**/53.4 | 28.8/- | 27.4/- | 32.6/- | 22.0/- | **40.1**/61.0 |
| mal_Mlym | 16.3/- | 14.2/- | 34.6/- | 34.1/- | **38.5**/51.6 | 31.7/- | 30.4/- | 27.4/- | 21.1/- | **41.9**/62.9 |
| mar_Deva | 16.1/- | 14.7/- | **36.1**/- | 32.7/- | 31.7/48.0 | 30.8/- | 30.0/- | 19.8/- | 18.3/- | **42.8**/63.8 |
| ory_Orya | 13.9/- | 10.1/- | 31.7/- | 31.0/- | **34.6**/45.7 | 30.1/- | 28.6/- | 24.4/- | 20.9/- | **44.3**/64.4 |
| pan_Guru | 26.9/- | 21.9/- | 39.0/- | 35.1/- | **39.8**/49.0 | 35.8/- | 34.2/- | 27.0/- | 28.5/- | **47.1**/66.3 |
| tam_Taml | 16.3/- | 14.9/- | 31.9/- | 29.8/- | **36.8**/53.7 | 28.6/- | 27.7/- | 28.0/- | 20.0/- | **39.5**/60.8 |
| tel_Telu | 22.0/- | 20.4/- | 38.8/- | 37.3/- | **43.2**/55.9 | 33.5/- | 32.7/- | 30.6/- | 30.5/- | **46.1**/65.5 |

Table 32: **Comparison on FLORES-101 devtest on Indian Languages.** We report spBLEU for all and chrF++ where available, and bold the best score. **(a)** IndicBART (Dabre et al., 2021), **(b)** IndicTrans (Ramesh et al., 2022), **(c)** Google Translate, **(d)** Microsoft Translate. Numbers for **(c)** and **(d)** are taken from (Ramesh et al., 2022). NLLB-200 outperforms other translation systems on most of the directions. Improvements are more significant in the xx-eng_Latn directions.

HUAWEI

# Evaluation / Results (2)

- Comparison with academic systems
- Comparison
  - **BLEU** / chrF++

| | eng-xx | | xx-eng | |
|---|---|---|---|---|
| | Published | NLLB-200 | Published | NLLB-200 |
| ces | [b]**26.5**/- | 25.2/50.6 | [d]**35.3**/- | 33.6/56.8 |
| deu | [a]**44.9**/- | 33.0/59.2 | [a]**42.6**/- | 37.7/60.5 |
| est | [a]26.5/- | **27.0**/55.7 | [a]**38.6**/- | 34.7/59.1 |
| fin | [a]**32.1**/- | 27.7/57.7 | [a]**40.5**/- | 28.8/53.7 |
| fra | [a]**46.7**/- | 44.2/65.7 | [a]**43.9**/- | 41.9/63.9 |
| guj | [d]**17.8**/- | 17.6/46.6 | [f]25.1/- | **31.2**/56.5 |
| hin | [f]25.5/- | **26.0**/51.5 | [f]29.7/- | **37.4**/61.9 |
| kaz | [i]15.5/- | **34.8**/61.5 | [i]**30.5**/- | 30.2/56.0 |
| lit | [a]17.0/- | **37.0**/63.9 | [a]**36.8**/- | 29.7/56.4 |
| lvs | [a]**25.0**/- | 24.8/50.8 | [a]**28.6**/- | 24.8/50.8 |
| ron | [a]41.2/- | **41.5**/58.0 | [h]**43.8**/- | 43.4/64.7 |
| rus | [a]31.7/- | **44.8**/65.1 | [a]39.8/- | **39.9**/61.9 |
| spa | [e]33.5/- | **37.2**/59.3 | [e]34.5/- | **37.6**/59.9 |
| tur | [a]**32.7**/- | 23.3/54.2 | [a]**35.0**/- | 34.3/58.3 |
| zho | [b]**35.1**/- | 33.9/22.7 | [a]**28.9**/- | 28.5/53.9 |

(a) WMT

| | eng-xx | | xx-eng | |
|---|---|---|---|---|
| | Published | NLLB-200 | Published | NLLB-200 |
| arb | [b]22.0/- | **25**/47.2 | [b]44.5/- | **44.7**/63.7 |
| deu | [k]25.5/- | **31.6**/57.8 | [k]28.0/- | **36.5**/57.5 |
| fra | [g]40.0/- | **43.0**/65.6 | [g]39.4/- | **45.8**/64.8 |
| ita | [b]38.1/- | **42.5**/64.4 | [b]43.3/- | **48.2**/66.5 |
| jpn | [c]19.4/- | **19.5**/21.5 | [c]19.1/- | **22.6**/46.1 |
| kor | [c]**22.6**/- | 22.5/27.9 | [c]24.6/- | **25.4**/48.0 |
| nld | [c]34.8/- | **34.9**/60.2 | [c]**43.3**/- | 41.0/60.9 |
| pes | [j]06.5/- | **15.5**/39.2 | [j]18.4/- | **42.3**/61.3 |
| pol | [j]16.1/- | **21.1**/48.3 | [j]18.3/- | **27.1**/48.2 |
| ron | [k]25.2/- | **29.4**/55.5 | [k]31.8/- | **42.0**/62.0 |
| rus | [j]11.2/- | **24.0**/47.0 | [j]19.3/- | **30.1**/51.3 |
| vie | [c]**35.4**/- | 34.8/53.7 | [c]36.1/- | **36.6**/57.1 |

(b) IWSLT

HUAWEI

# Evaluation / Results (3)

- Human Evaluation
  - 15 from Eng
  - 15 non-Eng
  - 4 to Eng

- XSTS score
  - From 1 to 5
  - Smart aggregation

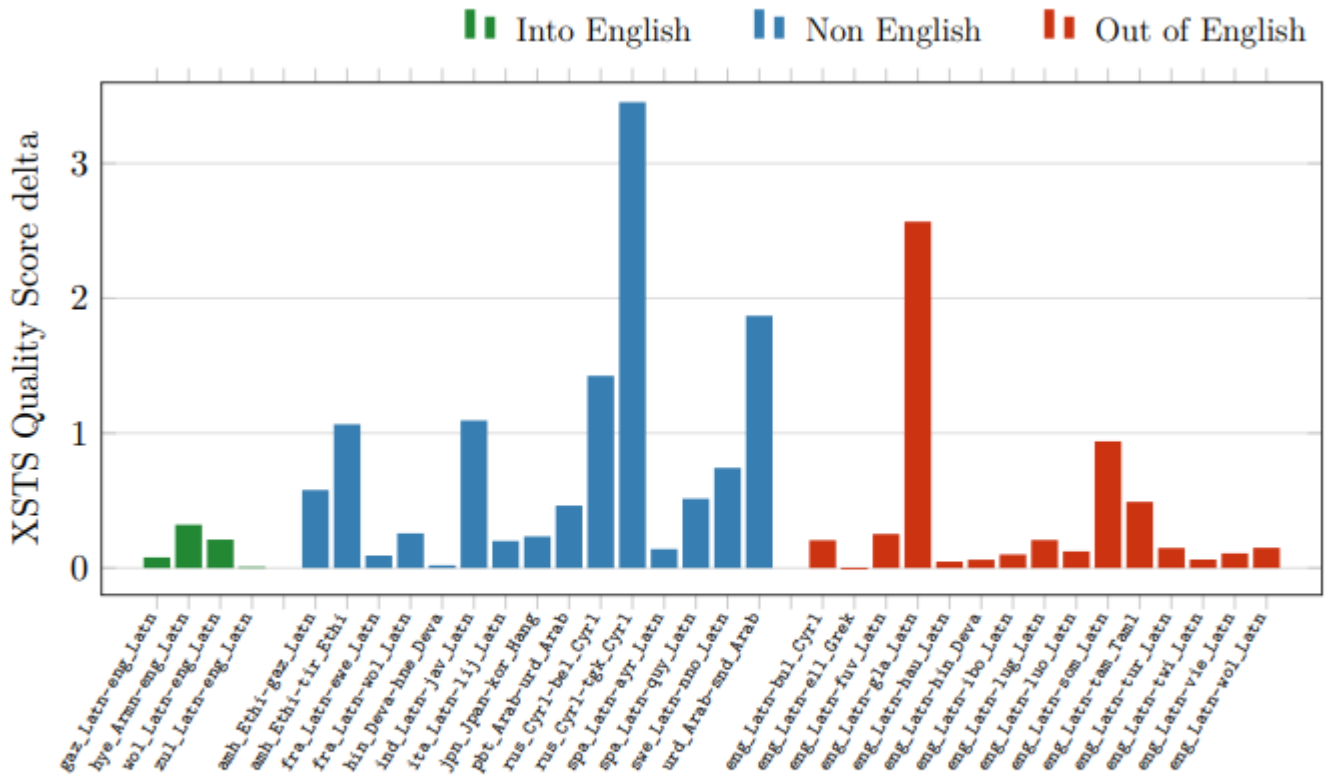



Figure 37: **XSTS Human Evaluation Quality Delta.** Delta between NLLB-200 model and dense baseline. We show average improvements 0.53 across all tested directions. Average improvement for into English directions of 0.15, out of English 0.36 and, average over non-English directions 0.81

HUAWEI

# Thank you.

Bring digital to every person, home, and organization for a fully connected, intelligent world.

Valentin Malykh, PhD
Valentin.Malykh@Huawei.com
http://val.maly.hk