No Language Left Behind (NLLB) Scaling Human-Centered Machine Translation

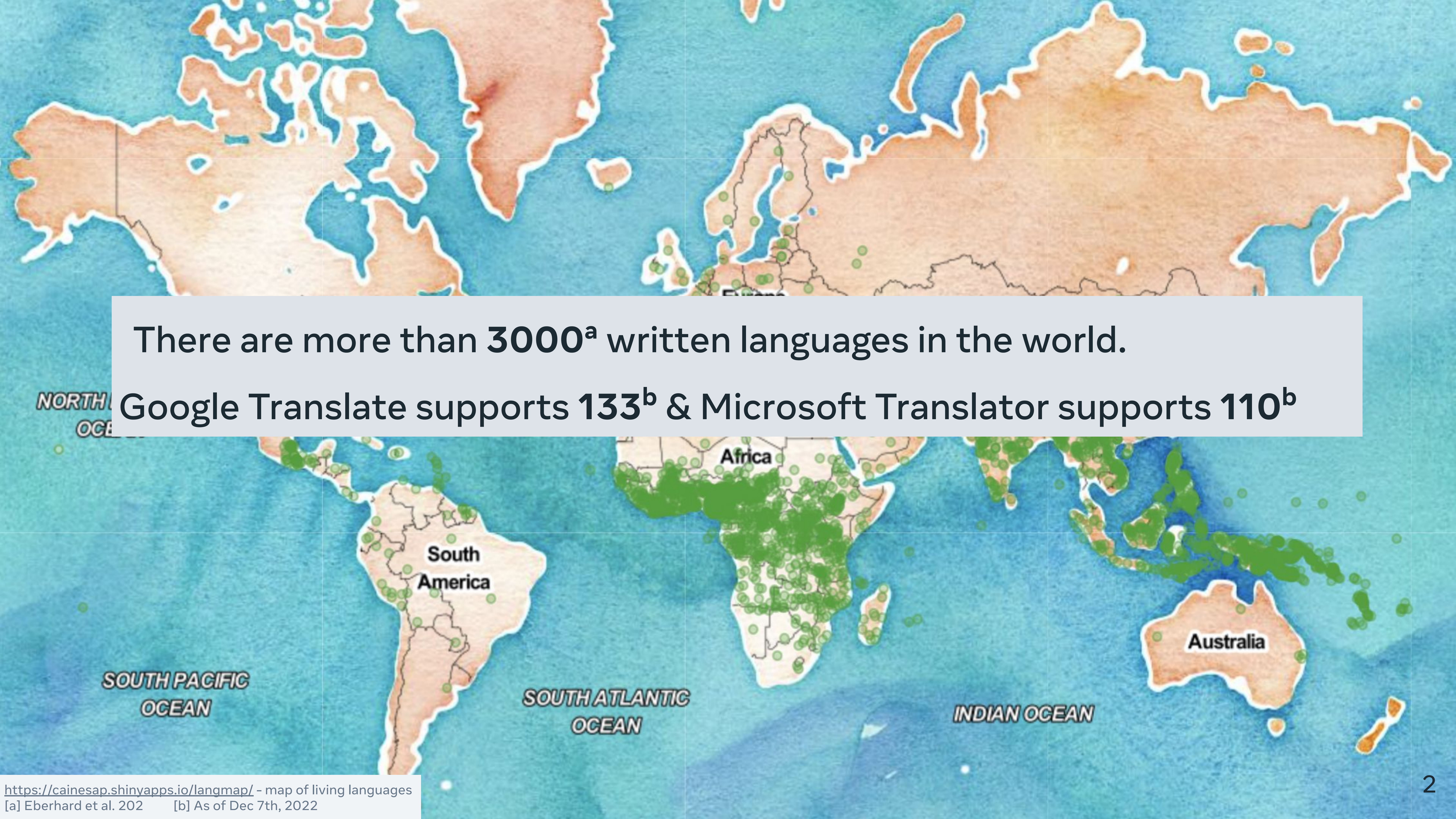
NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, John Hoffman, Semarley Jarrett, Kaushik Ram Sadagopan, Dirk Rowe, Shannon Spruit, Chau Tran, Pierre Andrews, Necip Fazil Ayan, Shruti Bhosale, Sergey Edunov, Angela Fan, Cynthia Gao, Vedanuj Goswami, Francisco Guzmán, Philipp Koehn, Alexandre Mourachko, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, Jeff Wang

Maha Elbayad
Research Scientist, FAIR

Meta AI

There are more than **3000**[a] written languages in the world.

Google Translate supports **133**[b] & Microsoft Translator supports **110**[b]



https://cainesap.shinyapps.io/langmap/ - map of living languages

[a] Eberhard et al. 202 [b] As of Dec 7th, 2022

NORTH STAR

Develop a general-purpose **universal** machine translation model capable of translating between **any two** languages in various domains.

- The majority of improvements in MT are for high-resource languages.
- Handling low-resource, underserved languages brings additional challenges:
  - Creating training data
  - Training multilingual MT models
  - Properly evaluating performance

THE NLLB EFFORT

How we structured our project to take on these challenges?

Multilingual Machine Translation is a multi-faceted problem.

Our research effort is taken on by an interdisciplinary team:

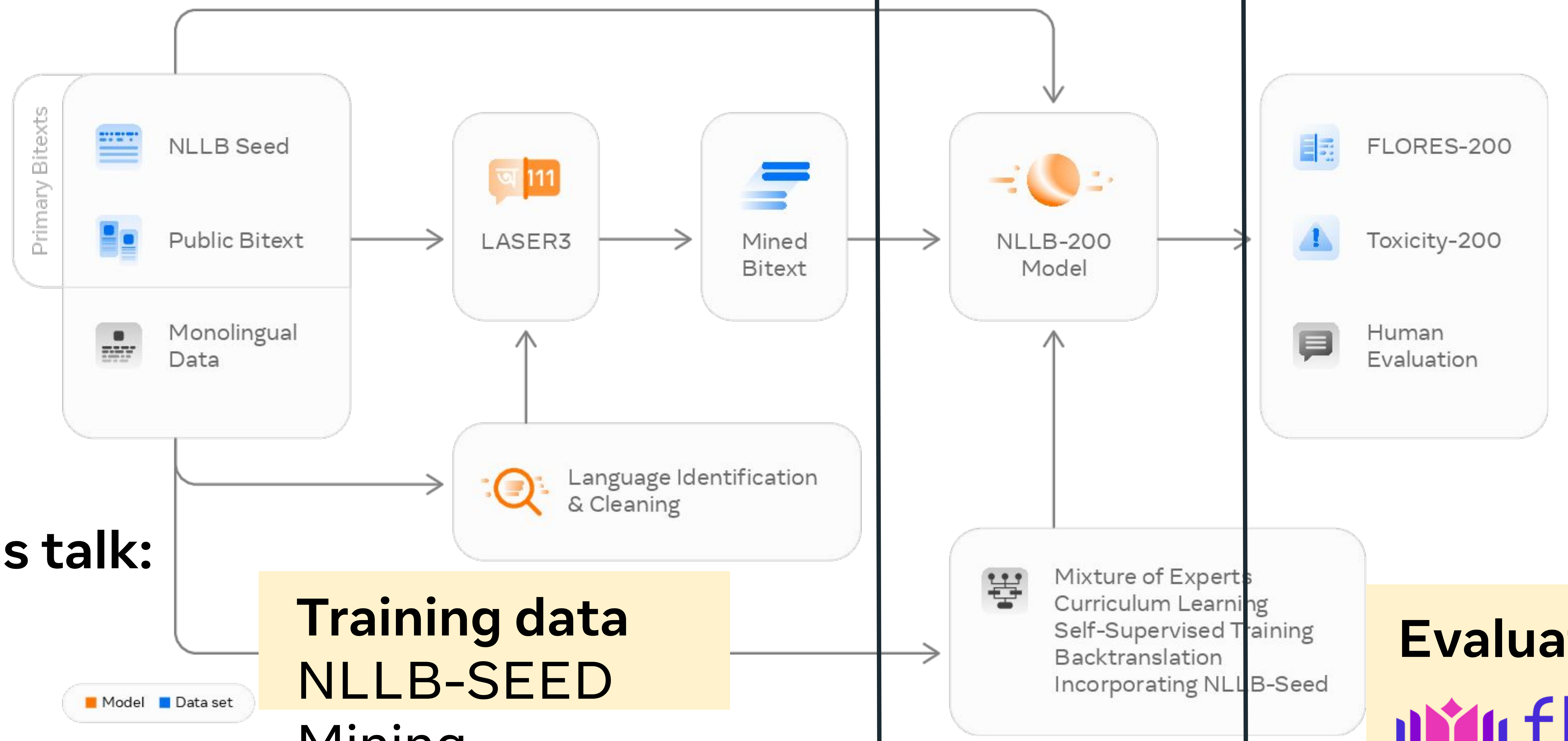
- Humanities i.e., Philosophy, Ethics
- Social scientific i.e., Sociology, Linguistics
- Technical i.e., Computer Science, Statistics

THE NLLB EFFORT

Our team was structured around our key challenges

	Data	Modeling	Evaluation
Research Question	How can we collect enough training data for low-resource languages?	How can we scale multilingual MT to 200 languages?	How can we evaluate across 200 languages with confidence and mitigate toxicity in the model outputs?
Deliverables	High quality aligned sentences covering 200 languages	Final MT model with optimum architecture and training strategy	High quality evaluation benchmark Toxicity lists covering 200 languages

THE NLLB EFFORT



In this talk:

Training data
NLLB-SEED
Mining
Back-translation

Modeling

Evaluation data
flores

Data

1. Multilingual Benchmark Dataset (FLORES-200)
2. Bitext Seed Data (NLLB-SEED)

Data

1. FLORES-200 (Benchmark)

A high-quality evaluation dataset or a reliable benchmark can help assess progress. The ability to evaluate allows us to compare different approaches and understand what requires further research and development.

- High quality, **many-to-many** benchmark dataset.
- The same 3,001 sentences in 204 languages (> 40,000 directions).
- English source collected from Wikinews, Wikijunior, Wikivoyage.
- Translated and reviewed by professional translators and reviewers.
- Focus on **low resource languages.**

flores

Data

2. NLLB-SEED

Human-translated bitext data in 39 low-resource languages to train models that require parallel data

Purpose:

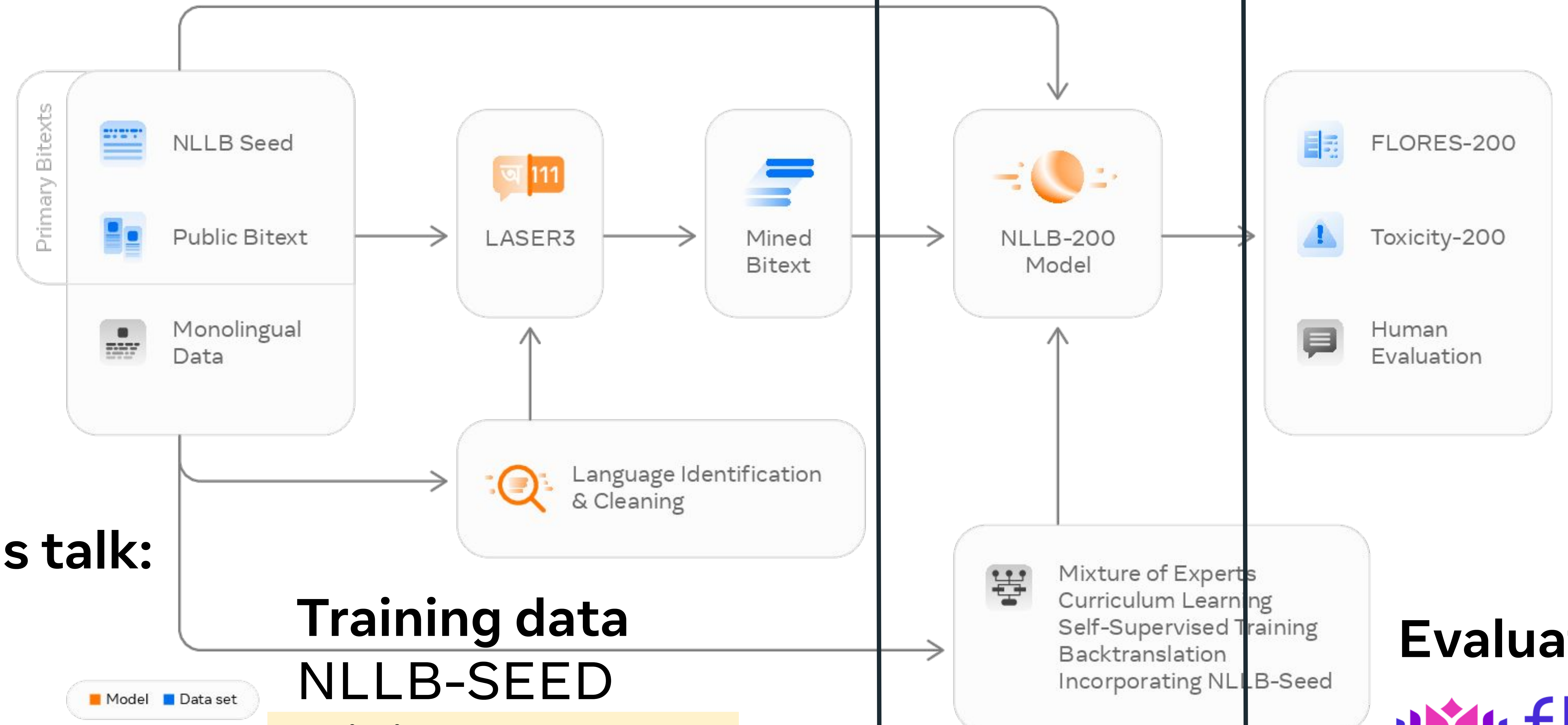
- Supporting language identification for new languages
- Aligned bitext to help train translation models
- Domain finetuning (Ex: adapting general-purpose translation models to the Wikipedia domain)

Data Collection Process:

- Sampled from Wikimedia's *List of articles every Wikipedia should have*[1]
- Sampled triplets of continuous sentences from English Wikipedia articles in 11 categories incl. People, History, Philosophy and Religion, Geography, etc.

[1] https://meta.wikimedia.org/wiki/List_of_articles_every_Wikipedia_should_have/Expanded

THE NLLB EFFORT



In this talk:

Training data

NLLB-SEED

Mining

Back-translation

Modeling

Evaluation data

flores

Modeling

1. Bitext Mining
2. Back-translation
3. Training large models

Modeling

1. Bitext Mining

We extend existing datasets with large-scale data mining (Schwenk et al. 2021) i.e., collecting non-aligned monolingual data and identifying sentences that have a high probability of being translations of each other.

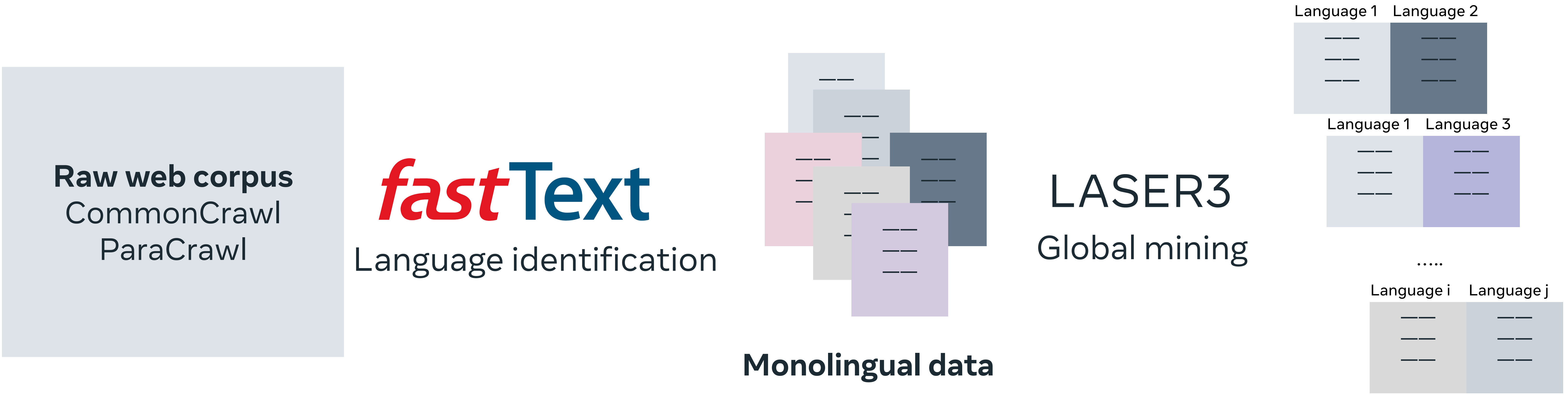
Modeling

1. Bitext Mining

There are two components to the data mining pipeline:

a. **Language IDentification (LID)** systems to predict the primary language for a span of text – **FastText** (Grave et al. 2018)

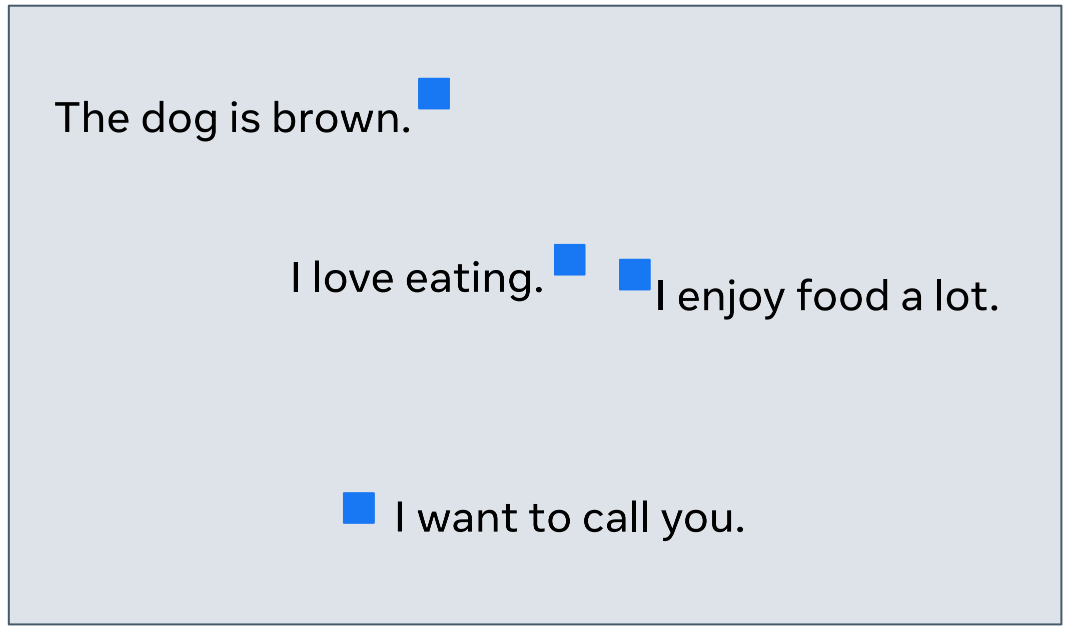
b. **Multilingual Sentence Encoders** to embed sentences and find similar semantically similar sentences in different languages – **LASER3** (Heffernan et al. 2022)



Modeling

1. Bitext Mining

b. **Multilingual Sentence Encoders** to embed sentences and find semantically similar ones in different languages – LASER (Artexte and Schwenk, 2019), LaBSE (Feng et al, 2020).



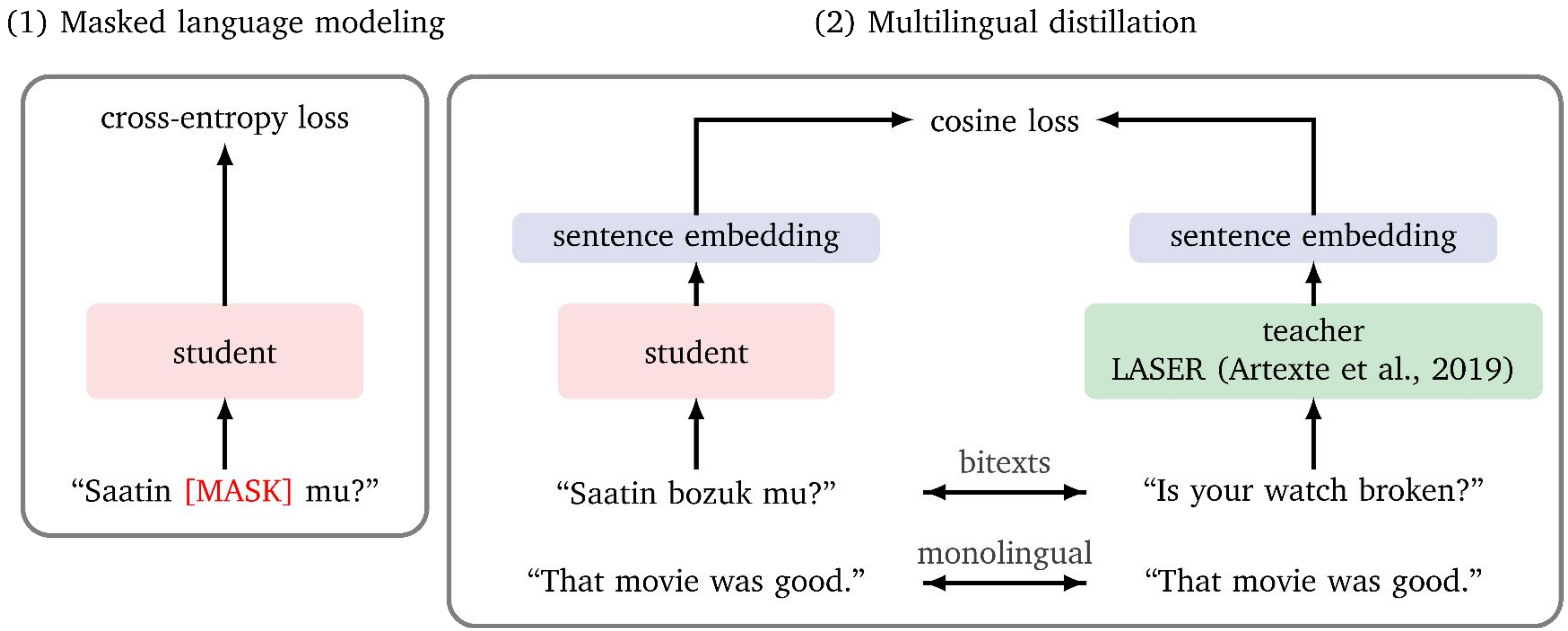
Sentences with similar meaning are ***close***.

Sentences with similar meaning are ***close* independently of their language**

Modeling

1. Bitext Mining

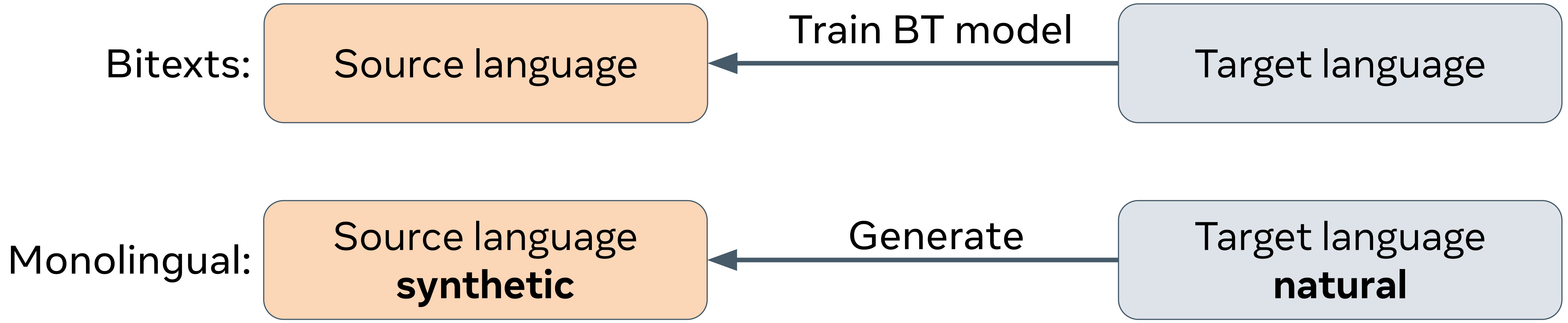
b. **Multilingual Sentence Encoders** LASER3 encoders are trained independently via distillation (Heffernan et al. 2022)



Modeling

2. Back-translation

Create parallel corpora noisy on the source side via machine translation (Sennrich et al. 2016; Edunov et al. 2018).



We generate BT data with two models:

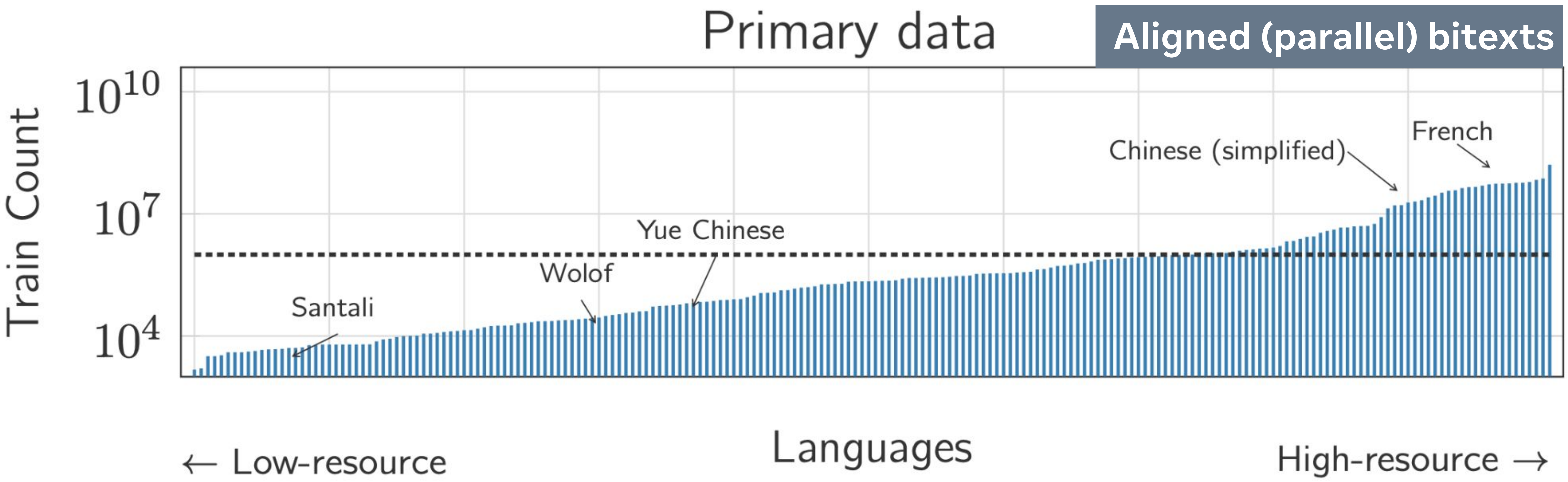
- **MmtBT**, a multilingual neural MT model.
- **SmtBT**, a series of bilingual MOSES models.

Modeling

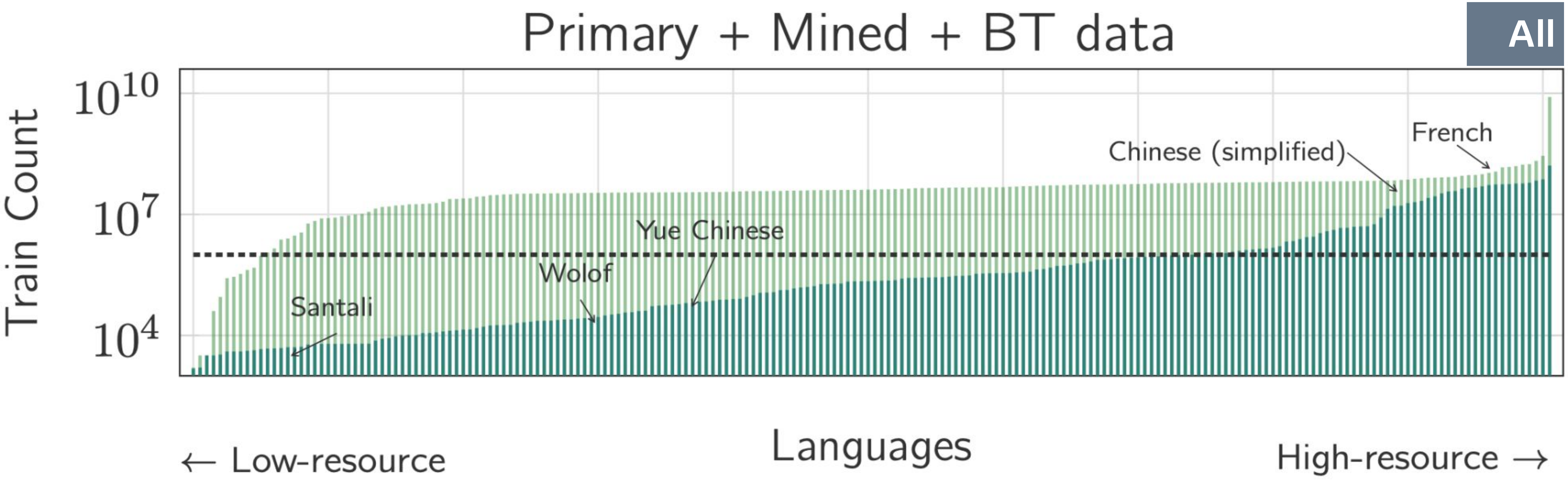
Summary- sources of training data

Source	Human Aligned?	Noisy?	Limited Size?	Model-Dependent?	Models Used
NLLB-Seed	✓	✗	✓	✗	—
PublicBitext	✗	✓	✓	✗	—
Mined	✗	✓	✗	✓	Sentence Encoders
MmtBT	✗	✓	✗	✓	Multilingual
SmtBT	✗	✓	✗	✓	Bilingual MOSES
Ideal Data	✓	✗	✗	✗	—

Modeling


Primary data
Aligned (parallel) bitexts
Train Count
10^{10}
10^7
10^4
French
Chinese (simplified)
Yue Chinese
Wolof
Santali
Languages
← Low-resource
High-resource →

Modeling

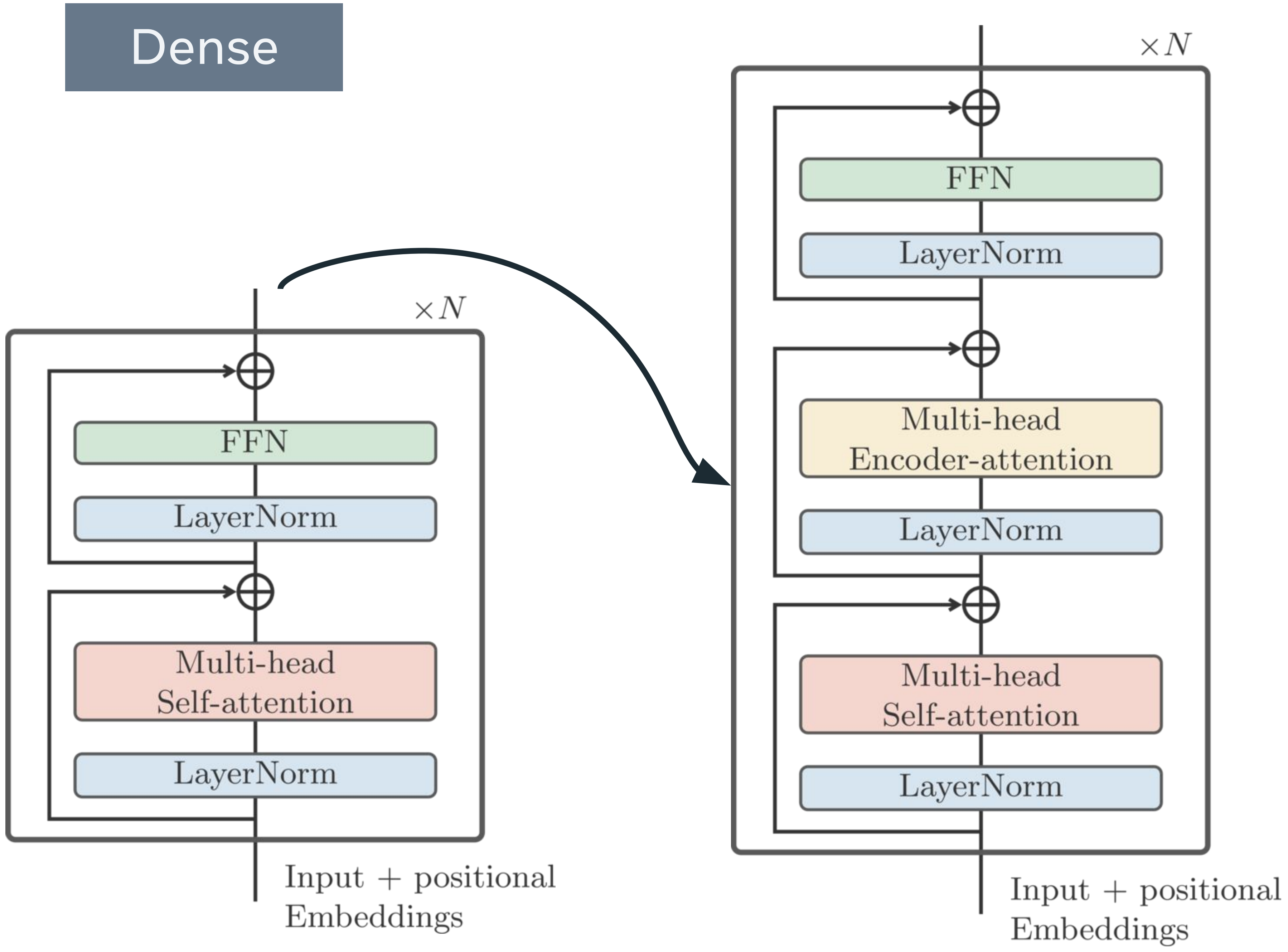


Primary + Mined + BT data
All
Train Count
10^{10}
10^{7}
10^{4}
French
Chinese (simplified)
Yue Chinese
Wolof
Santali
Languages
← Low-resource
High-resource →

Modeling

3. Training large models - Mixture of Experts

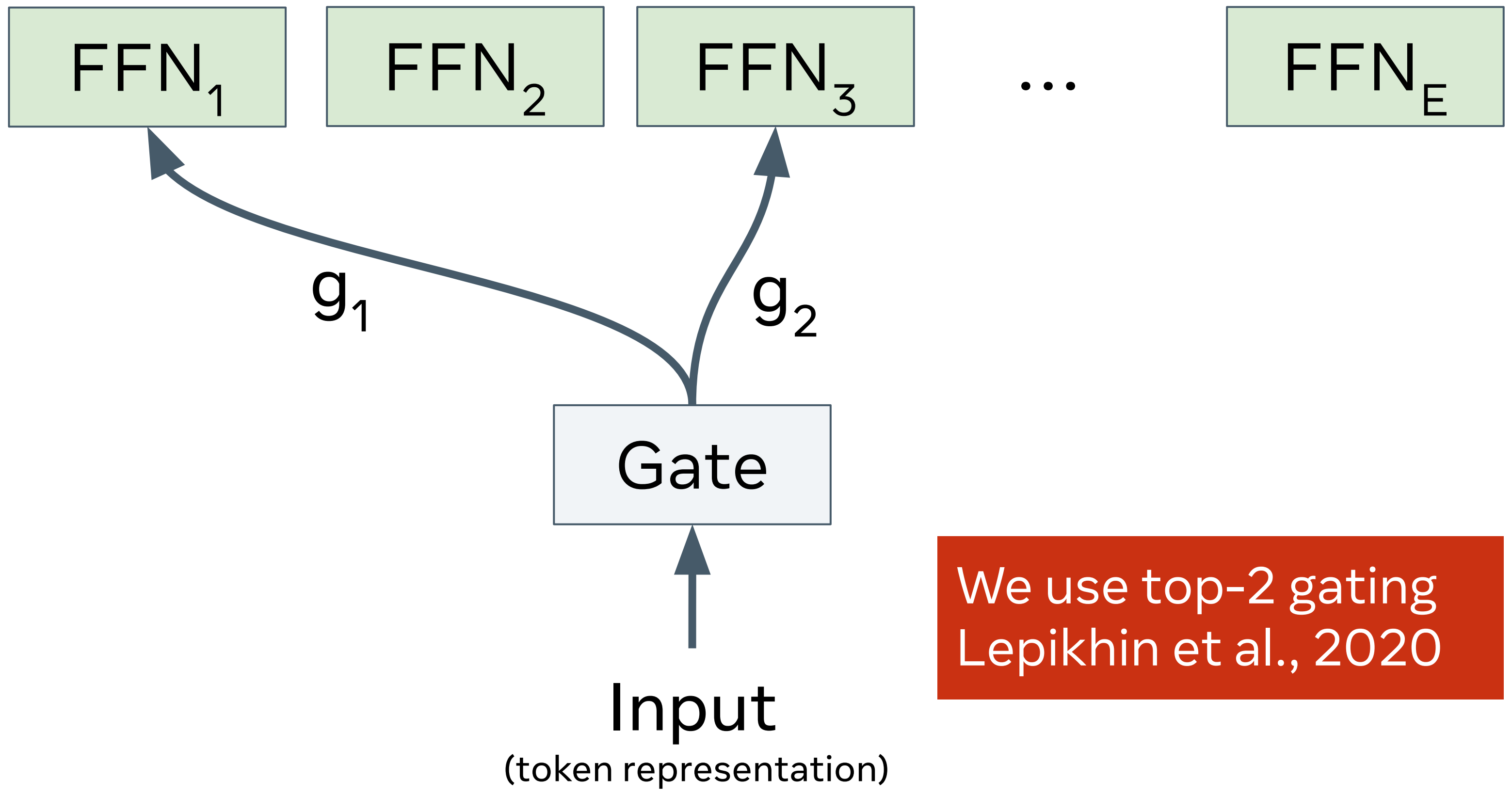
Dense



Encoder
Source sentence prefixed with <source_language>

Decoder
Target sentence prefixed with <target_language>

Sparsely Gated Mixture of Experts (MoE)



We use top-2 gating
Lepikhin et al., 2020

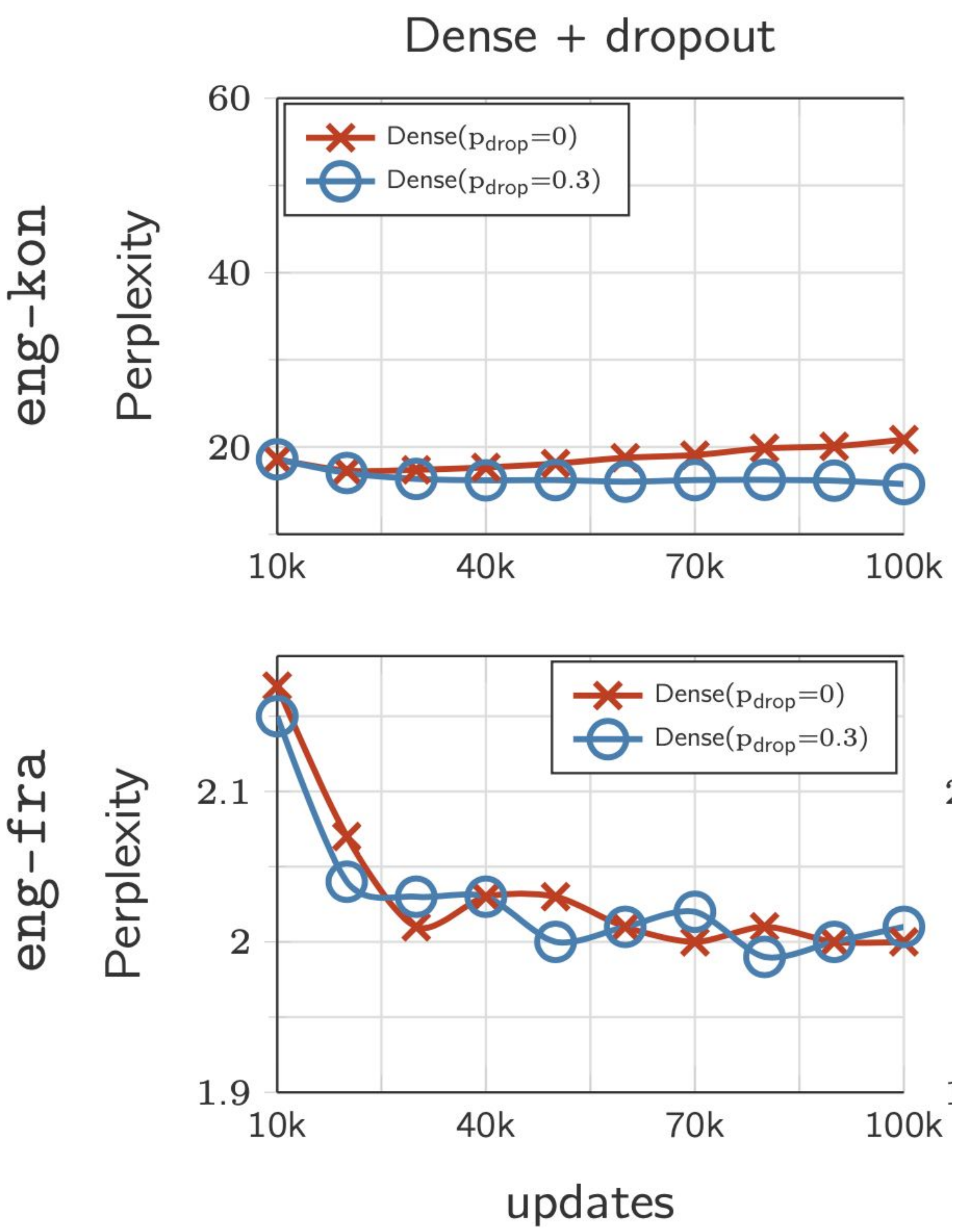
Replace every other FFN in the Transformer model with an MoE FFN layer

Modeling

3. Training large models - the issue of overfitting low-resource languages

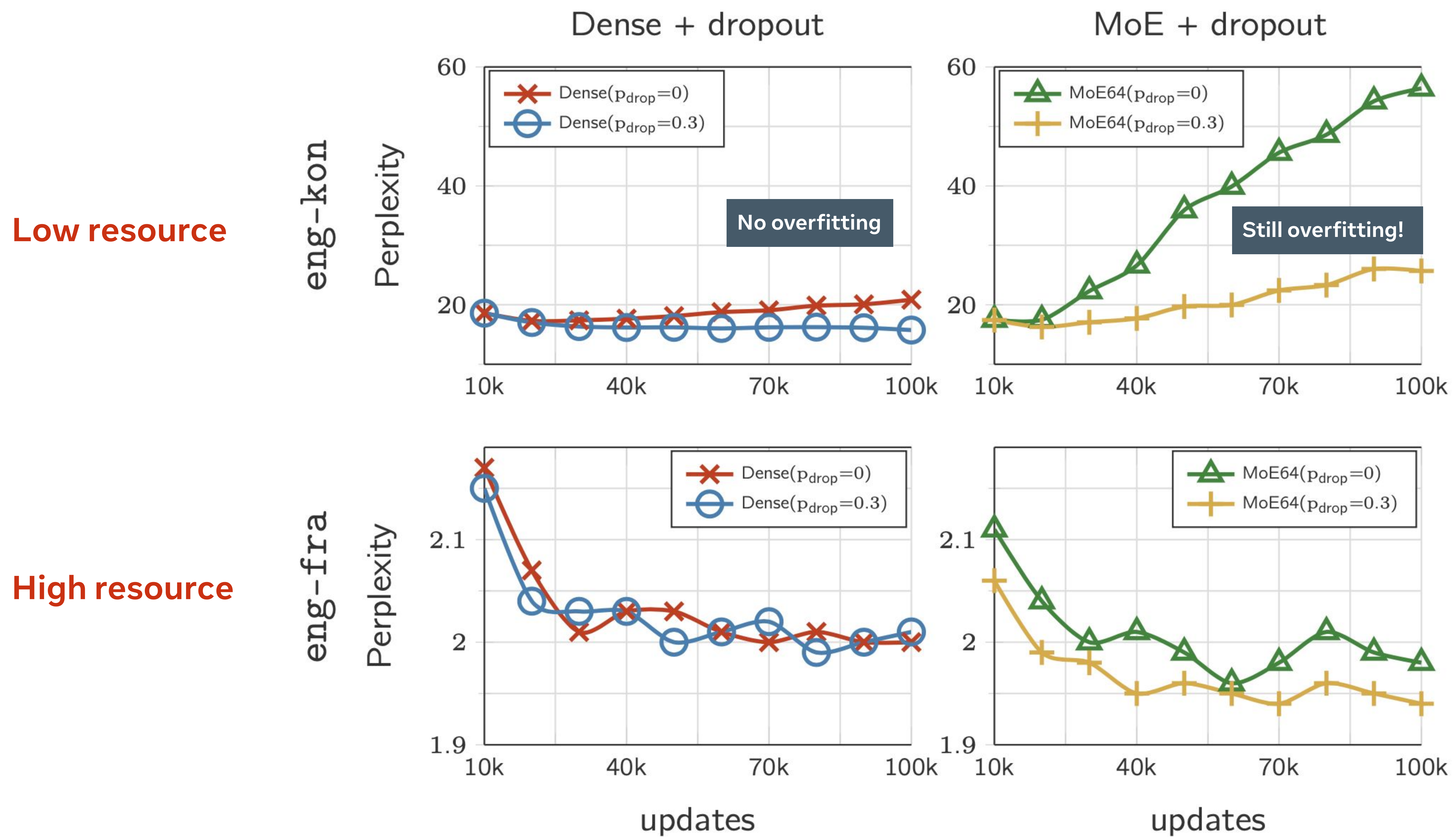
Low resource

High resource



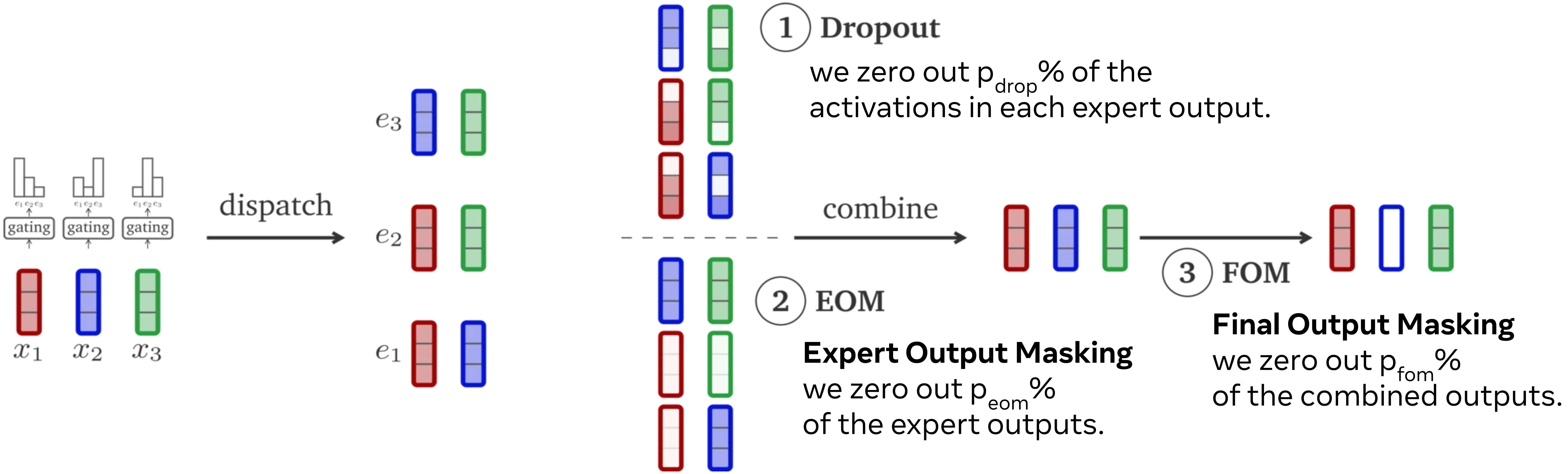
Modeling

3. Training large models - the issue of overfitting low-resource languages



Modeling

3. Training large models - Addressing overfitting



We combine these methods with **Curriculum learning**, where we introduce translation directions that overfit early, later in the training process.

Results

Modeling

Results - seed datasets

Experimental setup. We train small bilingual models on 8 directions, we first train on the small amounts of pre-existing publicly available parallel data (primary) and then adding seed datasets

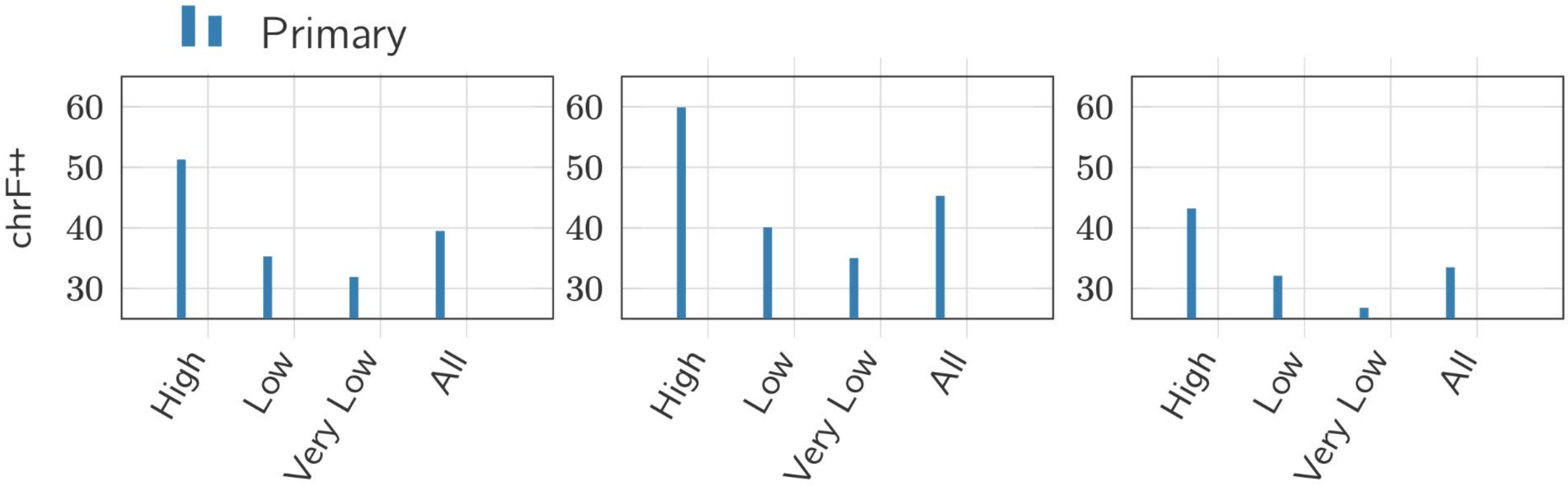
Back-translation, as well as a number of other augmentation approaches, rely on the presence of a "**seed model**" to bootstrap the system.

	public bitext		seed data		combined
	#data	chrF++	#data	chrF++	chrF++
ban-eng	10.2k	13.1	6.2k	**20.8**	22.2
eng-ban		15.9		**20.6**	21.9
dik-eng	16.9k	12.9	6.2k	**16.1**	17.0
eng-dik		9.0		**13.7**	13.1
fuv-eng	12.1k	15.6	6.2k	**16.3**	18.1
eng-fuv		9.2		**9.8**	13.5
mri-eng	31.3k	16.7	6.2k	**17.4**	26.8
eng-mir		23.2		**24.3**	31.5

Modeling

Results

Experimental setup. We train dense 3.3B Transformer encoder-decoder models with model dimension 2048, FFN dimension 8192, 16 attention heads and 48 layers (24 encoder, 24 decoder) for these data ablation experiments.

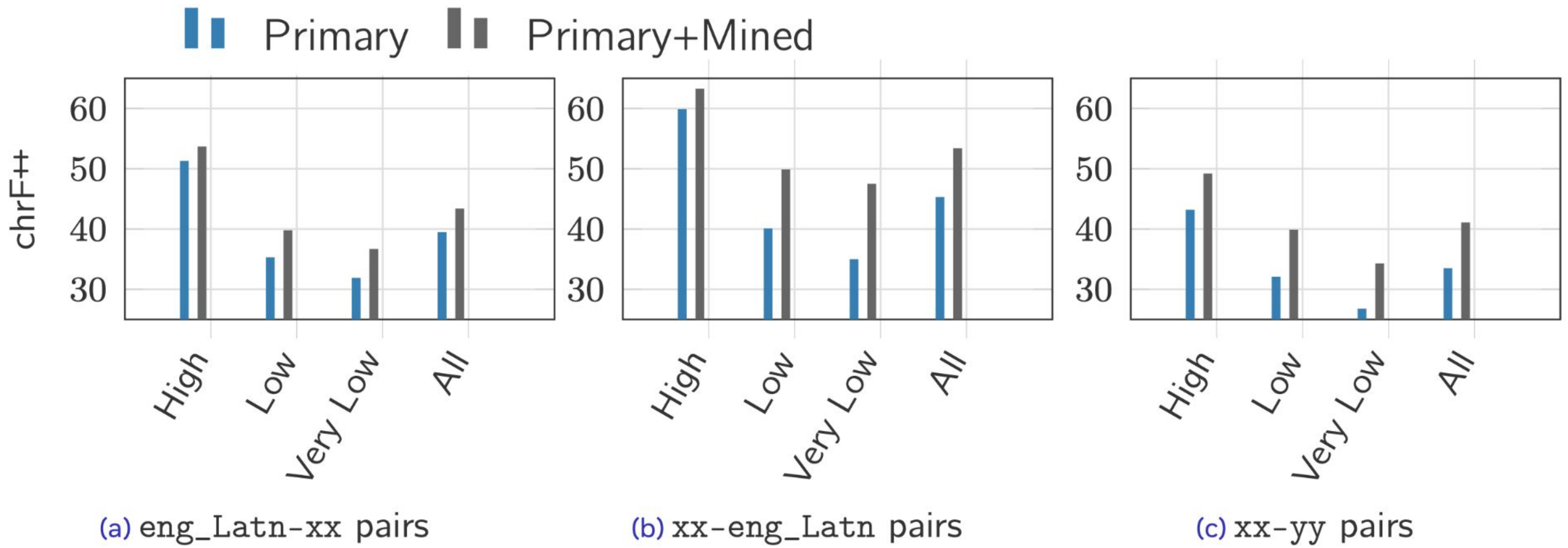


(a) eng_Latn-xx pairs (b) xx-eng_Latn pairs (c) xx-yy pairs

Modeling

Results

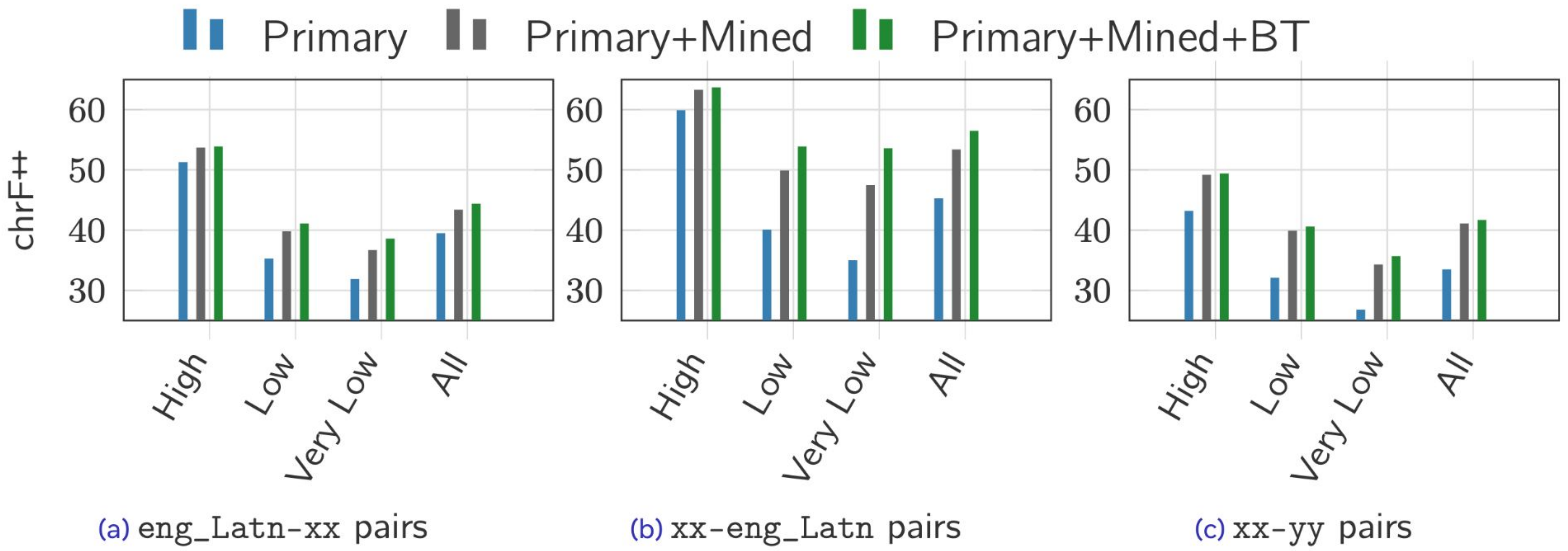
Experimental setup. We train dense 3.3B Transformer encoder-decoder models with model dimension 2048, FFN dimension 8192, 16 attention heads and 48 layers (24 encoder, 24 decoder) for these data ablation experiments.



Modeling

Results

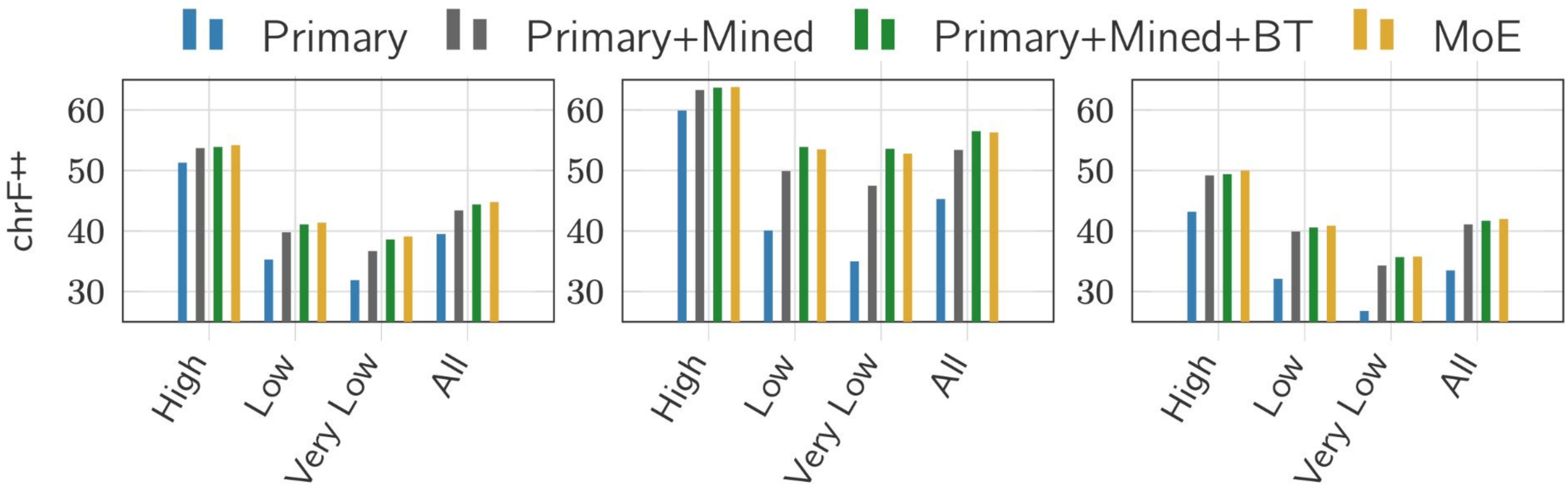
Experimental setup. We train dense 3.3B Transformer encoder-decoder models with model dimension 2048, FFN dimension 8192, 16 attention heads and 48 layers (24 encoder, 24 decoder) for these data ablation experiments.



Modeling

Results

Experimental setup. We train dense 3.3B Transformer encoder-decoder models with model dimension 2048, FFN dimension 8192, 16 attention heads and 48 layers (24 encoder, 24 decoder) for these data ablation experiments.

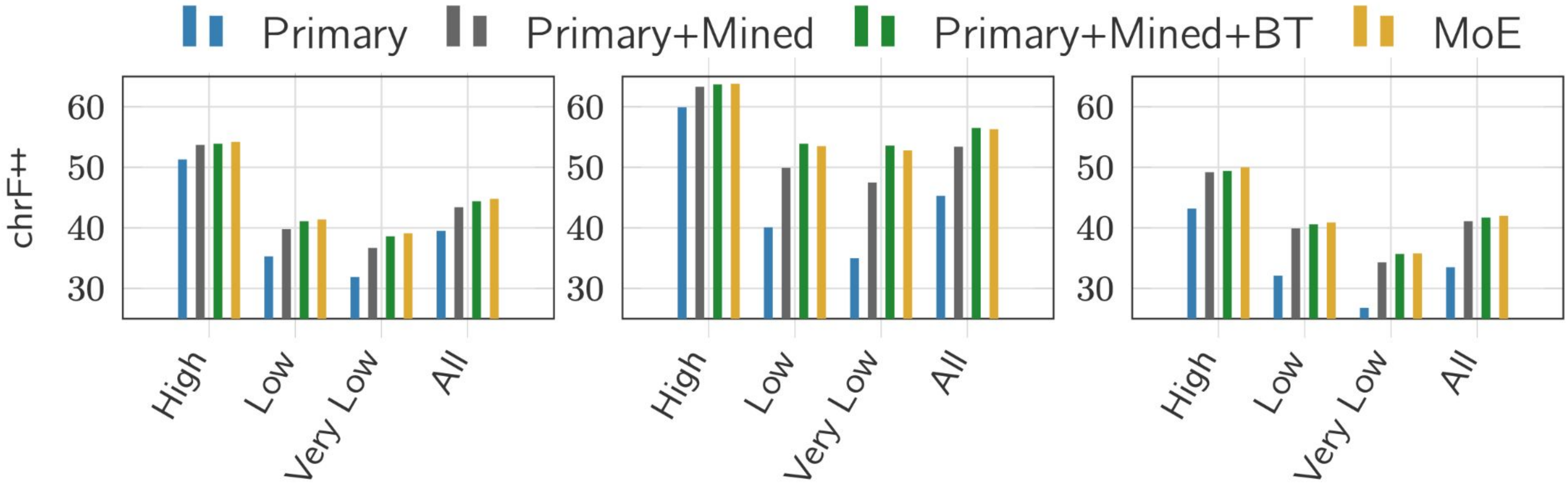


(a) eng_Latn-xx pairs (b) xx-eng_Latn pairs (c) xx-yy pairs

Modeling

Results

Experimental setup. We train dense 3.3B Transformer encoder-decoder models with model dimension 2048, FFN dimension 8192, 16 attention heads and 48 layers (24 encoder, 24 decoder) for these data ablation experiments.



(a) eng_Latn-xx pairs (b) xx-eng_Latn pairs (c) xx-yy pairs

Modeling

Results - NLLB-200 significantly outperforms previous SOTA.

Flores-101 devtest (spBLEU/chrF++)

	eng_Latn-xx	xx-eng_Latn	xx-yy	Avg.
87 languages				
M2M-100	-/-	-/-	-/-	13.6/-
Deepnet	-/-	-/-	-/-	18.6/-
NLLB-200	**35.4**/52.1	**42.4**/62.1	**25.2**/43.2	**25.5**/43.5
101 languages				
DeltaLM	26.6/-	33.2/-	16.4/-	16.7/-
NLLB-200	**34.0**/50.6	**41.2**/60.9	**23.7**/41.4	**24.0**/41.7

We also compare favorably to models trained on one language family (e.g. African languages with MMTAfrica and Mafand-MT or Indic languages with IndicBART and IndicTrans) - see tables 31 & 32 of the NLLB paper.

Modeling

Results - Performance on the new Flores-200

Flores-200 devtest (chrF++)

	eng_Latn-xx				xx-eng_Latn				xx-yy	Average
	all	high	low	v.low	all	high	low	v.low	all	all
chrF++	45.3	54.9	41.9	39.5	56.8	63.5	54.4	54.4	35.6	35.7
spBLEU	27.1	38.3	23.1	21.3	38.0	44.7	35.5	35.6	17.3	17.5

	xx-yy (supervised)				xx-yy (zero-shot)			
	all	high	low	v.low	all	high	low	v.low
chrF++	39.7	43.9	39.3	38.6	35.4	46.3	34.6	33.3
spBLEU	20.3	24.3	19.9	20.0	17.2	28.3	16.4	15.3

Flores-200 devtest - 102 Low-Resource Directions (spBLEU/chrF++)

	eng_Latn-xx		xx-eng_Latn		Average	
	low	v.low	low	v.low	low	v.low
Google Translate	**32.3/50.3**	**27.0/46.5**	35.9/57.1	35.8/57.0	34.1/53.7	31.3/51.7
NLLB-200	30.3/48.2	25.7/45.0	**41.3/60.4**	**41.1/60.3**	**35.8/54.3**	**33.4/52.6**

Modeling

Results - Out-of-domain generalization

Evaluation and comparison to state-of-the-art on sampled directions from WMT, IWSLT, WAT, Floresv1, TICO, Mafand, Autshumato and Madar. These benchmarks cover domains other than wikipedia (e.g., health, news, scripted talks, …)

	eng-xx		xx-eng	
	Published	NLLB-200	Published	NLLB-200
khm	[b] **5.9**/-	0.4/27.4	[b] 10.7/-	**16.8**/36.5
npi	[c] 7.4/-	**10.4**/39.0	[c] 14.5/-	**29.3**/54.8
pbt	[b] 9.3/-	**10.5**/34.3	[b] 15.7/-	**22.0**/46.8
sin	[c] 3.3/-	**11.6**/40.9	[c] 13.7/-	**23.7**/49.8

(a) Flores(v1)

	eng-xx		xx-eng	
	Published	NLLB-200	Published	NLLB-200
hin	[l] 22.1/-	**27.2**/51.5	[l] 32.9/-	**37.4**/61.9
khm	[l] 43.9/-	**45.8**/42.3	[l] 27.5/-	**39.1**/61.1
mya	[c] **39.2**/-	23.5/31.5	[c] **34.9**/-	32.7/57.9

(b) WAT

	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**/-	21.3/50.8	[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

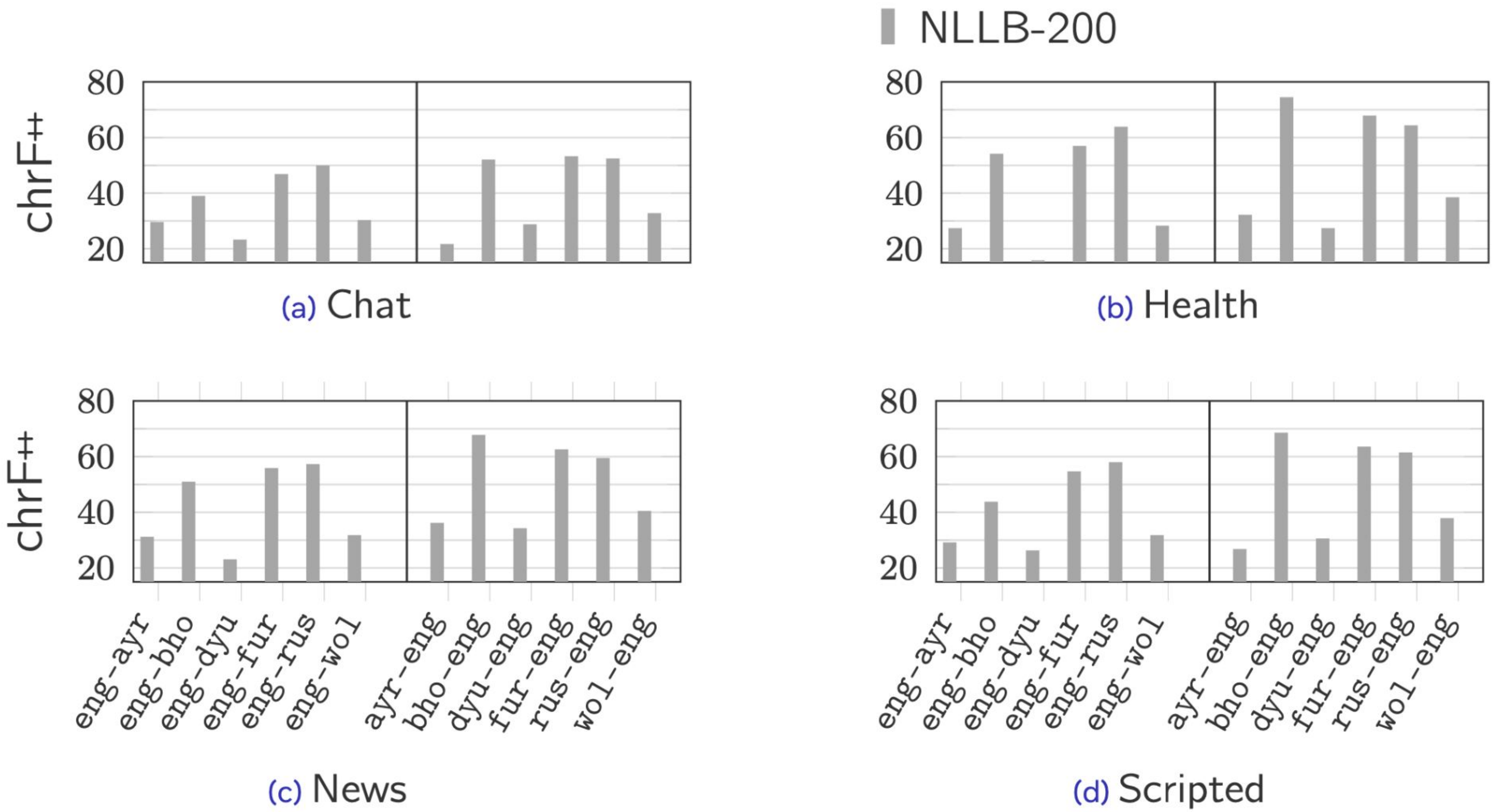
	eng-xx		xx-eng	
	Published	NLLB-200	Published	NLLB-200
arb	15.2/-	**34.1**/59.4	28.6/-	**49.6**/70.3
fra	37.6/-	**44.9**/64.4	39.4/-	**47.3**/65.4
gaz	0.6/-	**10.7**/44.0	2.1/-	**35.9**/57.2
hin	6.4/-	**46.2**/65.8	18.9/-	**58.0**/76.2
ind	41.3/-	**55.1**/74.8	34.9/-	**54.3**/73.5
lin	7.8/-	**24.6**/51.5	6.7/-	**33.7**/54.1
lug	3.0/-	**22.1**/48.6	5.6/-	**39.0**/58.2
mar	0.2/-	**16.1**/46.3	1.2/-	**44.3**/66.9
pes	8.5/-	**30.0**/55.6	15.1/-	**45.5**/67.5
por	47.3/-	**52.9**/72.9	48.6/-	**58.7**/76.5
rus	28.9/-	**35.7**/59.1	28.5/-	**41.2**/65.1
spa	48.7/-	**57.2**/74.9	46.8/-	**57.5**/75.9
swh	22.6/-	**34.1**/59.1	0.0/-	**49.6**/68.1
urd	2.8/-	**27.4**/53.3	0.0/-	**44.7**/66.9
zho	33.7/-	**42.0**/33.3	28.9/-	**37.6**/61.9
zsm	6.3/-	**52.4**/73.4	0.0/-	**58.8**/76.1
zul	11.7/-	**22.4**/55.1	25.5/-	**50.6**/68.4

	eng-xx		xx-eng	
	Adelani et al. (2022)	NLLB-200	Adelani et al. (2022)	NLLB-200
hau_Latn	**15.9**/**42.1**	8.2/34.8	**18.2**/**40.2**	13.5/37.9
ibo_Latn	**26.0**/**51.3**	23.9/50.4	**21.9**/**48.0**	**21.9**/46.1
lug_Latn	15.7/46.9	**25.8**/**55.2**	22.4/48.5	**30.9**/**54.4**
luo_Latn	12.0/39.4	**14.0**/**40.4**	14.3/38.3	**15.9**/**38.4**
swh_Latn	27.7/**57.2**	**30.7**/56.0	30.6/55.8	**39.3**/**60.8**
tsn_Latn	**31.9**/**59.5**	28.5/55.6	27.8/54.0	**37.3**/**60.2**
yor_Latn	13.9/**37.4**	**14.4**/36.3	18.0/41.0	**24.4**/**46.7**
zul_Latn	**22.9**/**56.3**	16.1/47.3	38.1/57.7	**40.3**/59.7
	fra-xx		xx-fra	
	Adelani et al. (2022)	NLLB-200	Adelani et al. (2022)	NLLB-200
bam_Latn	**24.7**/**49.9**	7.7/29.9	**25.8**/**49.0**	14.6/37.5
ewe_Latn	**8.9**/**37.5**	8.3/36.4	11.6/37.2	**19.4**/**42.6**
fon_Latn	**7.4**/**28.5**	3.4/21.8	**9.9**/**28.9**	8.9/28.7
mos_Latn	2.2/16.8	**5.4**/**27.6**	4.1/18.8	**6.1**/**23.5**
wol_Latn	**12.7**/**35.8**	9.1/29.9	**11.5**/**35.3**	9.5/30.2

Modeling

Results - Out-of-domain generalization with **Finetuning**

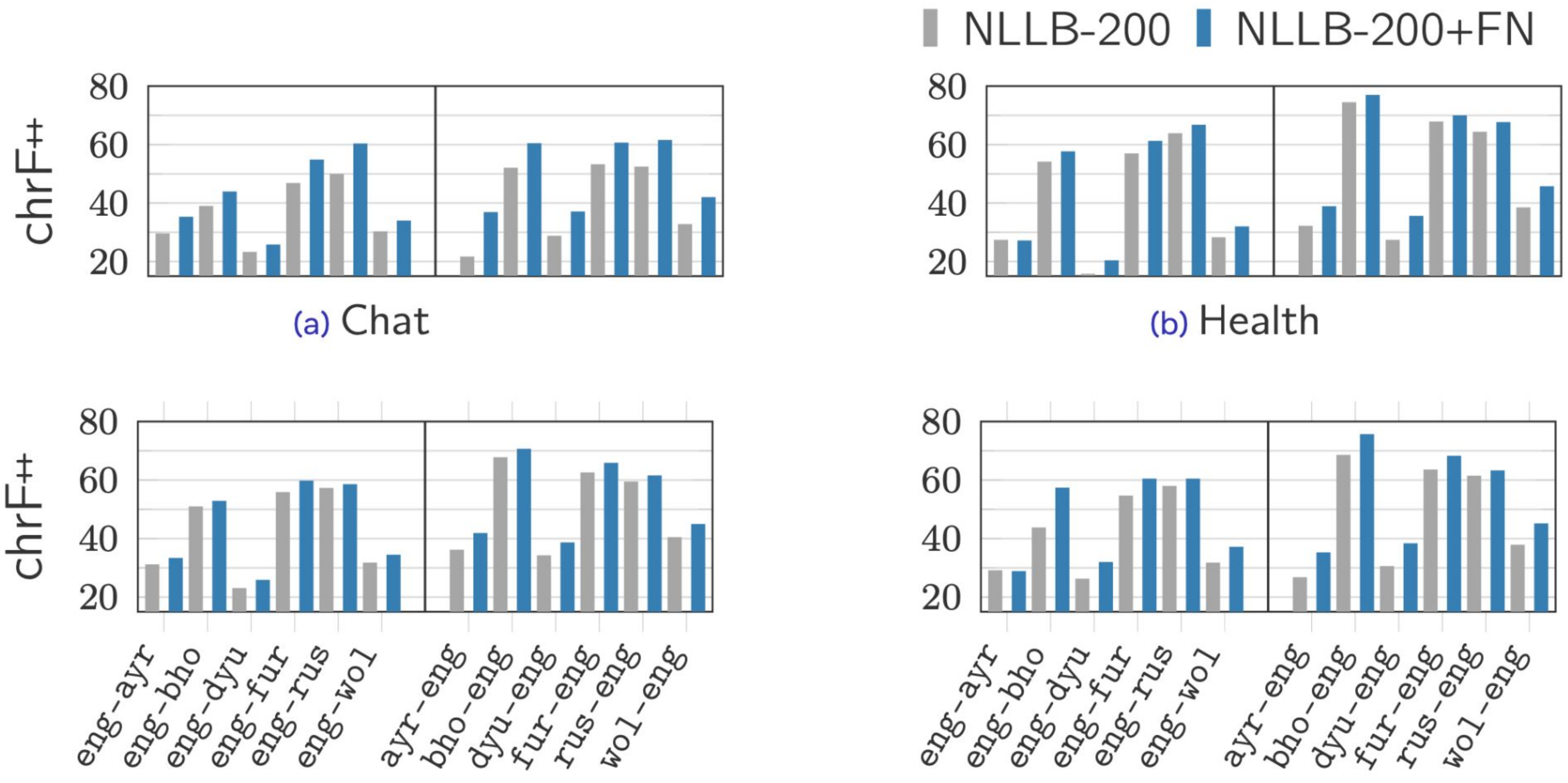
An additional dataset released, dubbed NLLB-MD (multi-domain) in 6 languages covering 3 domains (chat, news and health, ~~scripted~~).



Modeling

Results - Out-of-domain generalization with **Finetuning**

An additional dataset released, dubbed NLLB-MD (multi-domain) in 6 languages covering 3 domains (chat, news and health, ~~scripted~~).



(a) Chat (b) Health (c) News (d) Scripted

THE NLLB EFFORT

Project webpage: https://ai.facebook.com/research/no-language-left-behind/

The Paper: https://arxiv.org/abs/2207.04672

Demo with children stories https://nllb.metademolab.com/story/1

Codebases

Modeling: https://github.com/facebookresearch/fairseq/tree/nllb

LASER3 (sentence encoders): https://github.com/facebookresearch/LASER/blob/main/nllb

Stopes (data and mining pipelines): https://github.com/facebookresearch/stopes/

THE NLLB EFFORT

Models checkpoints

Final NMT models: https://github.com/facebookresearch/fairseq/tree/nllb#multilingual-translation-models

+ Different model sizes (1.3B, 3.3B and 54.5B) + distilled models (600M and 1.3B)
+ NLLB-200 translations, first and only instance of open sourcing model translations on such a large scale

LASER3 encoders: https://github.com/facebookresearch/LASER/blob/main/nllb

Data

Flores-200, NLLB-Seed, NLLB-MD, Toxicity-200: https://github.com/facebookresearch/flores

Mined bitexts: https://huggingface.co/datasets/allenai/nllb

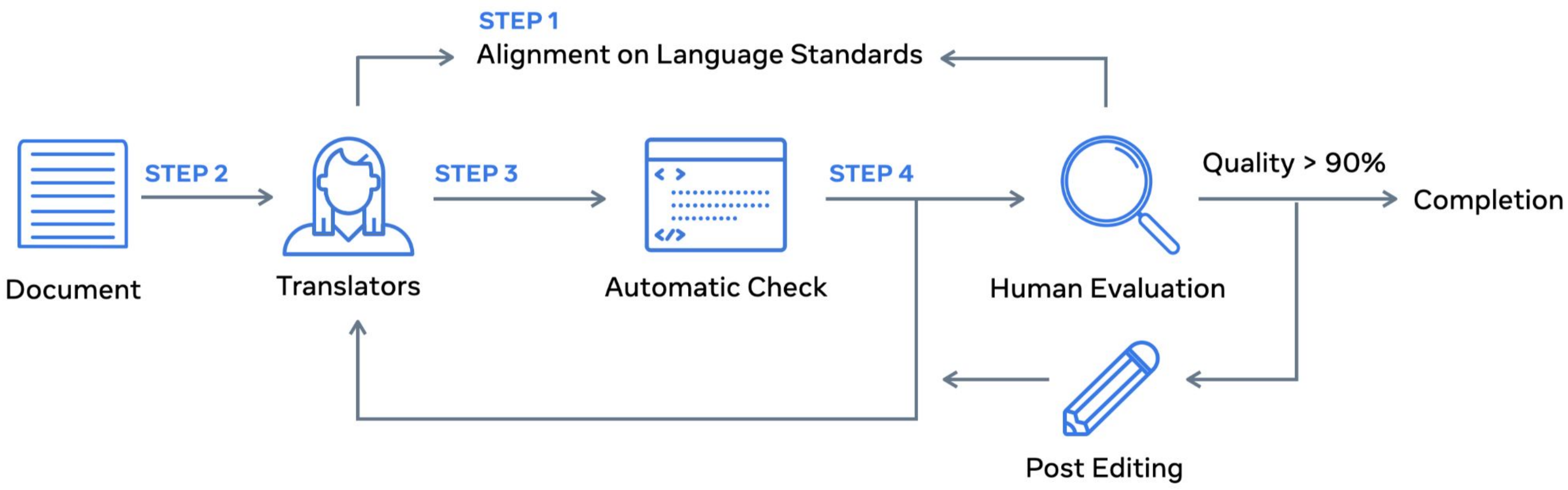
Meta AI

Data

1. Data Collection Processes: FLORES-200 (Benchmark)

Data Creation Process:

1. Translator + Reviewer Alignments
2. Initial Translation + QA + Arbitration
3. Full Translation
4. Automated and Linguistic Checks
5. Full QA by Third Party Reviewer
6. Arbitration (if applicable)
7. Rework + Spot Check (if applicable)
8. Final Delivery



flores

Data

2. Data Collection Challenges

Resourcing Challenges

- Difficulty in finding qualified resources for low-resource languages
- Finding and retaining resources
  - Consistency/continuity needed if working with new resources

Linguistic Challenges

- Dialectal Variations
- Lower levels of industry-wide standardization
  - Greater ambiguity
  - Higher subjectivity in assessing quality and consistent translations
- To tackle this:
  - Setting up alignments between translators and reviewers
  - Inevitable variations within an aligned dialect
    - How to balance preferential differences vs objective quality

Collection at Scale

- Language-specific challenges
- Long turnaround times
- Unexpected challenges throughout the whole process