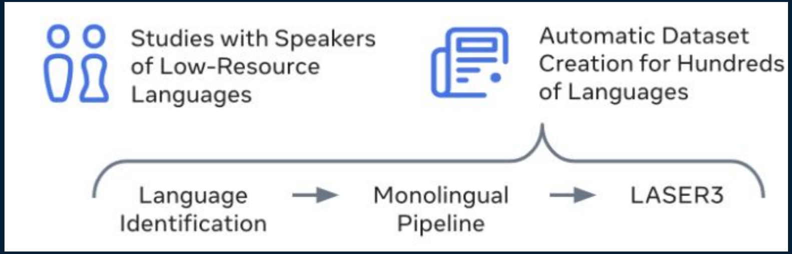
NLLB

No Language Left Behind

Hefty technical report!

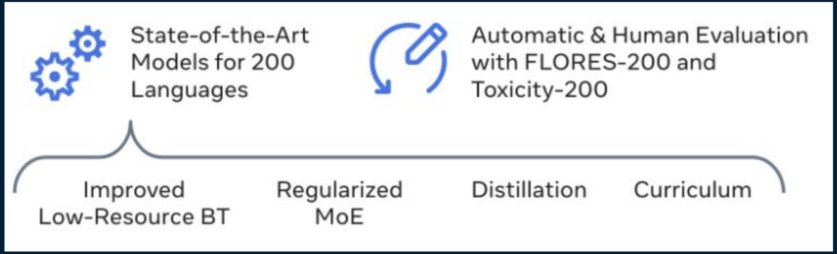
1 - Giulia

- Intro to Massively Multilingual Models
- NLLB ‘Human Centered’ approach
- ✨ Sprinkles of criticism ✨



2 - Ryan

- Mixture of Experts for massively multilingual translation



What is NLLB?

- Direct **translations** between **204 languages** (NLLB-200)

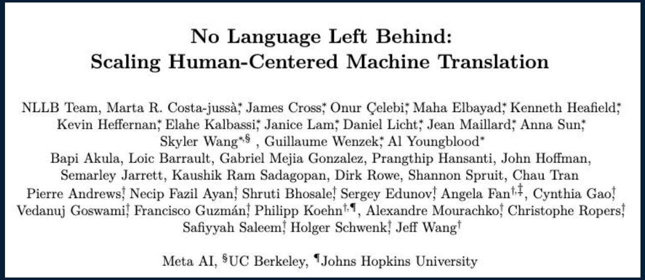
- *Machine translation* (MT) is the task of translating a sentence ***x*** in one language (**source language**) to a sentence ***y*** (**target language**)

    - **Early 1950s**: **rule-based systems** (didn't really work!)
    - **1990s-2010s**: **Statistical machine translation** (SMT)
    - **2010s-now**: **Neural machine translation** (NMT)


Article

Scaling neural machine translation to 200 languages

https://doi.org/10.1038/s41586-024-07335-x
NLLB Team*


No Language Left Behind: 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[†]

Meta AI, [§]UC Berkeley, [¶]Johns Hopkins University

Why Multilingual NLP?

- Most languages are 'Left-Behinds' (Joshi et al., 2019)
- 95% of languages in use today will never gain traction online (Kornai, 2013)

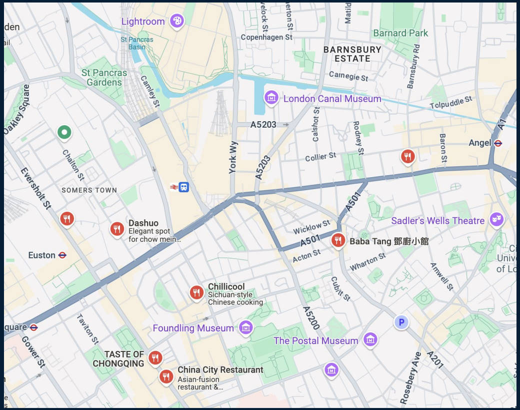
The limits of my world online are the limits of my world?

Class	5 Example Languages	#Langs	#Speakers	% of Total Langs
0	Dahalo, Warlpiri, Popoloca, Wallisian, Bora	2191	1.0B	88.17%
1	Cherokee, Fijian, Greenlandic, Bhojpuri, Navajo	222	1.0B	8.93%
2	Zulu, Konkani, Lao, Maltese, Irish	19	300M	0.76%
3	Indonesian, Ukranian, Cebuano, Afrikaans, Hebrew	28	1.1B	1.13%
4	Russian, Hungarian, Vietnamese, Dutch, Korean	18	1.6B	0.72%
5	English, Spanish, German, Japanese, French	7	2.5B	0.28%

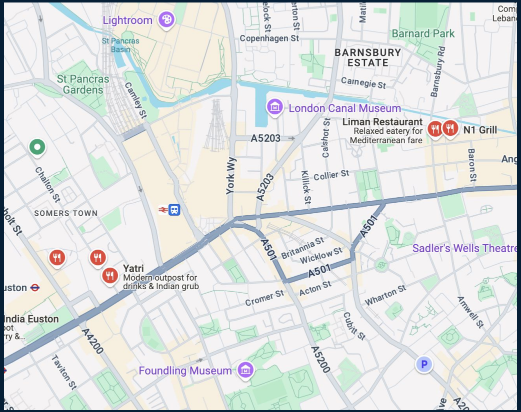
Why Multilingual NLP?

Inequality of information and representation can affect how we understand places, events, processes…

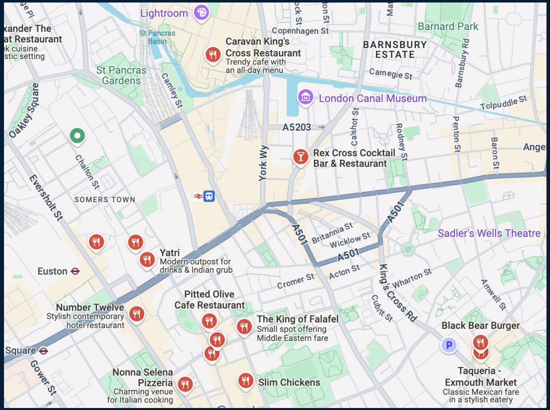
We're in King's Cross searching for…



…餐厅(ZH)



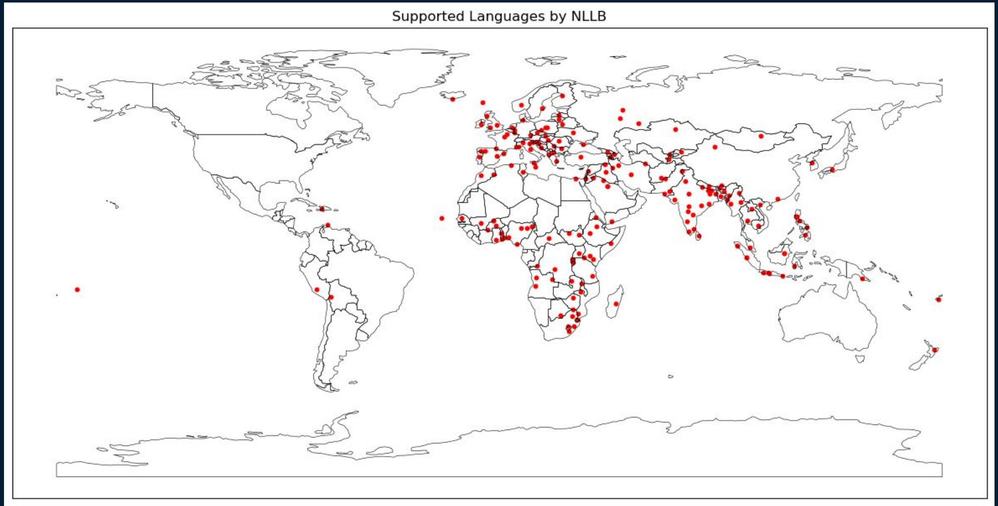
…রেস্টুরেন্ট(BGD)



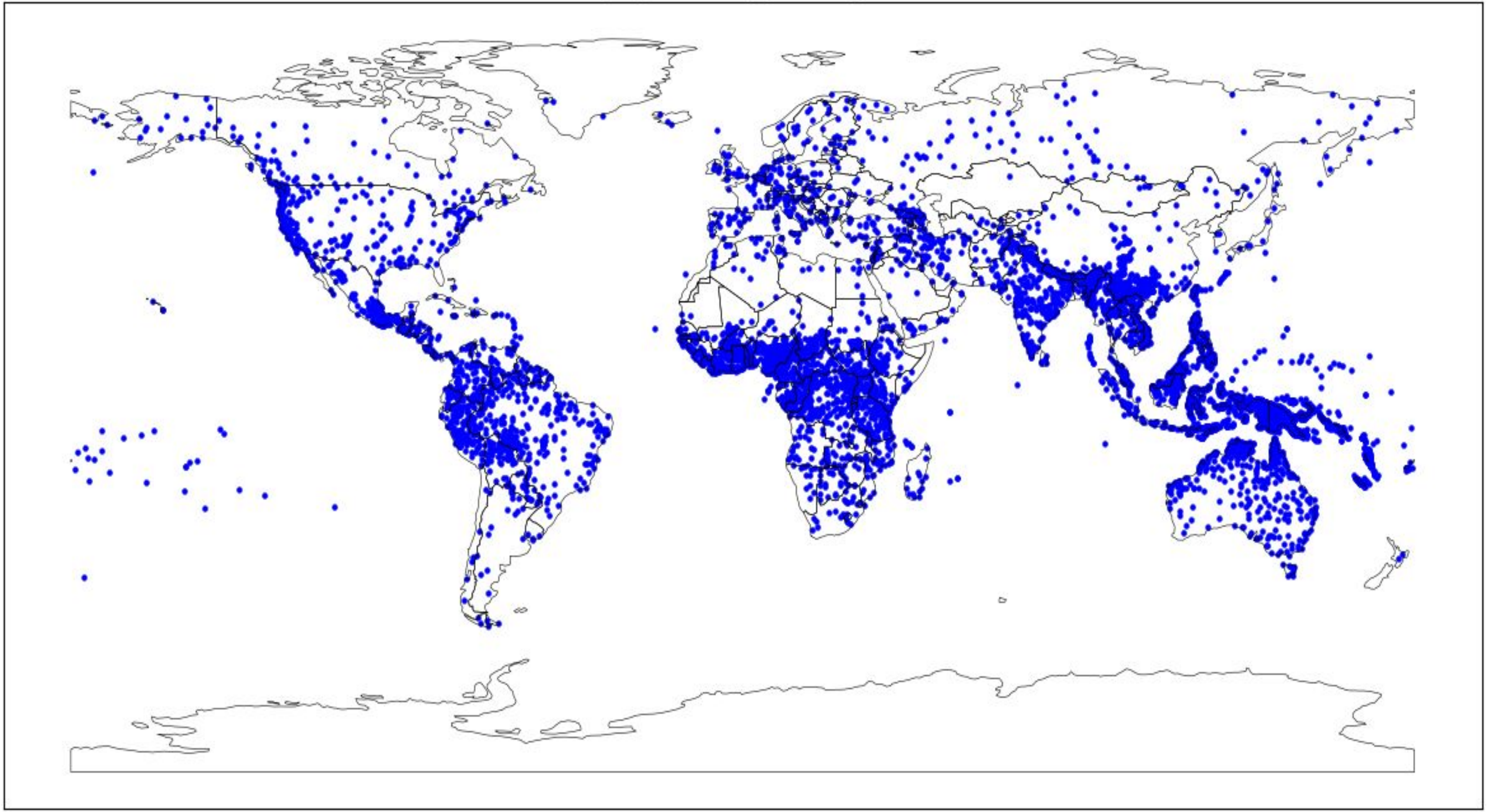
…restaurants(EN)

Why Multilingual NLP?

- Issues more pressing than restaurants in London…
  - Education 📚
  - Health ⚕️
  - Economic Empowerment 💰
- Giulia's 2 cents 🪙🪙:



Languages not supported by NLLB



NLLB ‘field research’

Discussion with native speakers to understand Users’ Needs:

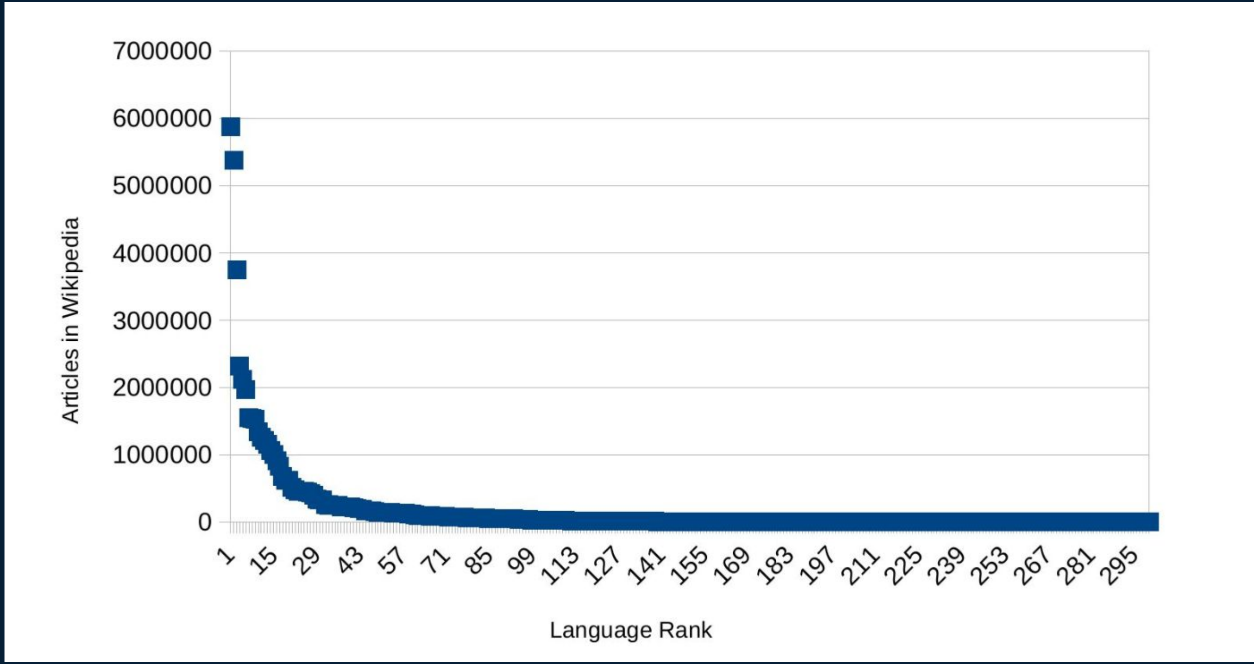
- **In the US**
- **Tech workers**
- **Igbo speakers example**
  - Igbo has 50 millions + speakers and is an official national language?

Key Challenges for low-resource NLP

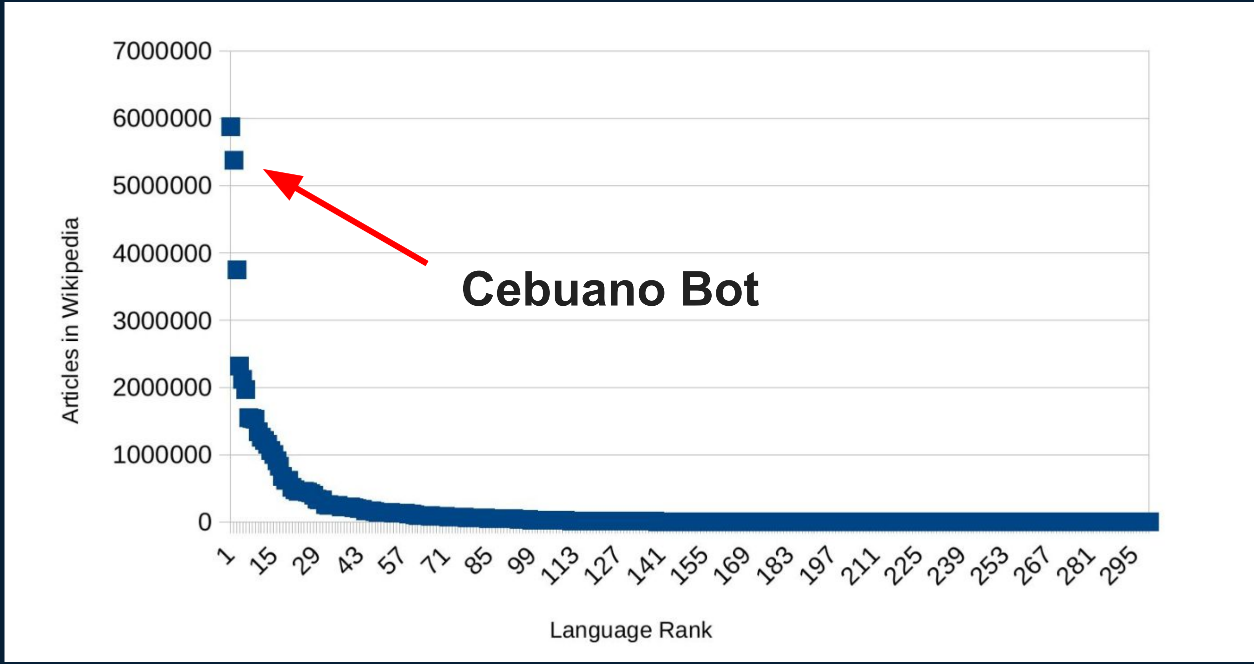
- In NLLB, a *low-resource language* is defined as those which had less than 1 million sentences of publicly available translations
    - Note: Some languages have no monolingual data online either...

- NMT models require large volumes of data to produce quality translations
    - Collecting data is expensive and logistically challenging

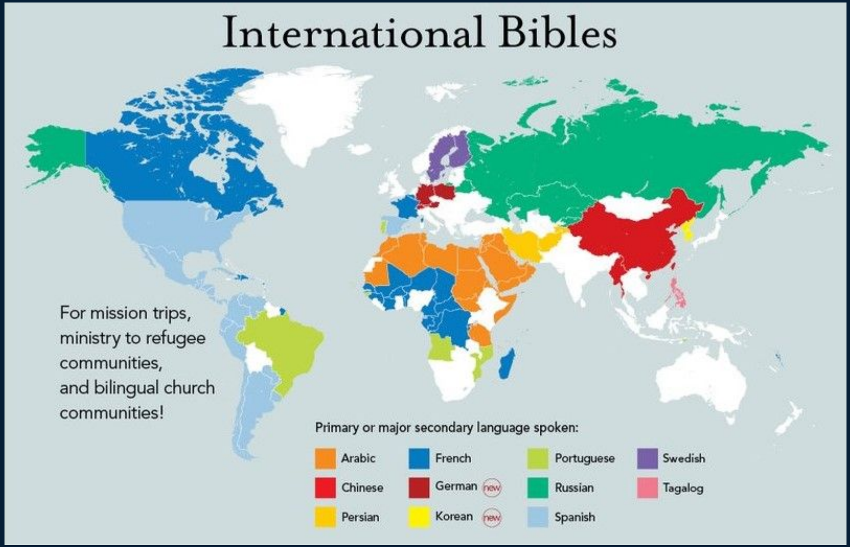
Key Challenges for low-resource NLP



Key Challenges for low-resource NLP



Solution: the Bible?



Languages with available Scriptures:

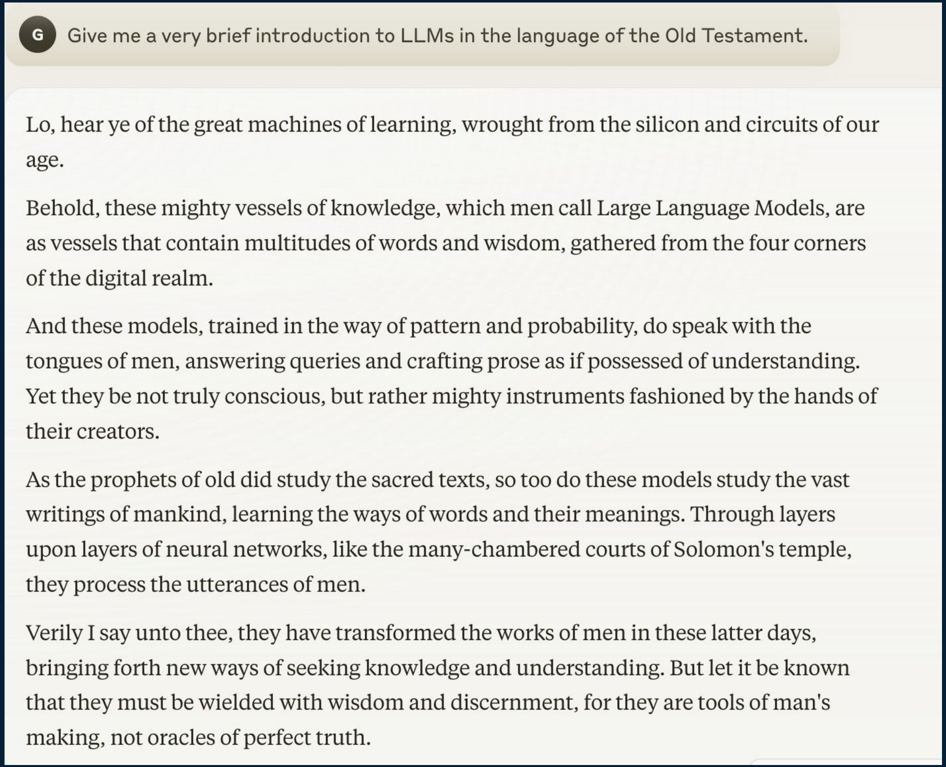
Full bible: 704

New Testament: 1551

Segment: 1.160

Total: 3.415!

Humans do not really speak like the bible…



G Give me a very brief introduction to LLMs in the language of the Old Testament.

Lo, hear ye of the great machines of learning, wrought from the silicon and circuits of our age.

Behold, these mighty vessels of knowledge, which men call Large Language Models, are as vessels that contain multitudes of words and wisdom, gathered from the four corners of the digital realm.

And these models, trained in the way of pattern and probability, do speak with the tongues of men, answering queries and crafting prose as if possessed of understanding. Yet they be not truly conscious, but rather mighty instruments fashioned by the hands of their creators.

As the prophets of old did study the sacred texts, so too do these models study the vast writings of mankind, learning the ways of words and their meanings. Through layers upon layers of neural networks, like the many-chambered courts of Solomon's temple, they process the utterances of men.

Verily I say unto thee, they have transformed the works of men in these latter days, bringing forth new ways of seeking knowledge and understanding. But let it be known that they must be wielded with wisdom and discernment, for they are tools of man's making, not oracles of perfect truth.

Joint Multilingual Learning in a Nutshell

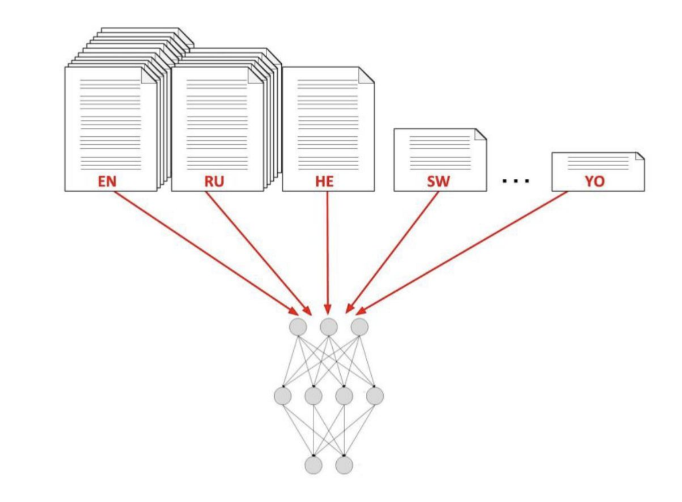


Image courtesy of *Yulia Tsvetkov*

Train a single model on a mix of datasets in all languages, to enable **data and parameter sharing** where possible

Cross-Lingual Transfer in a Nutshell

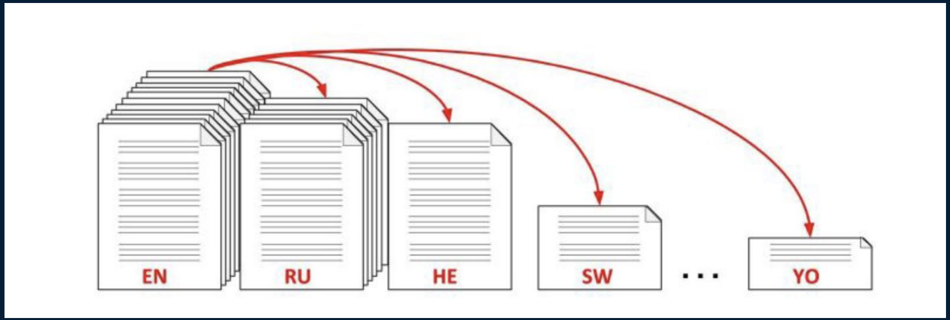
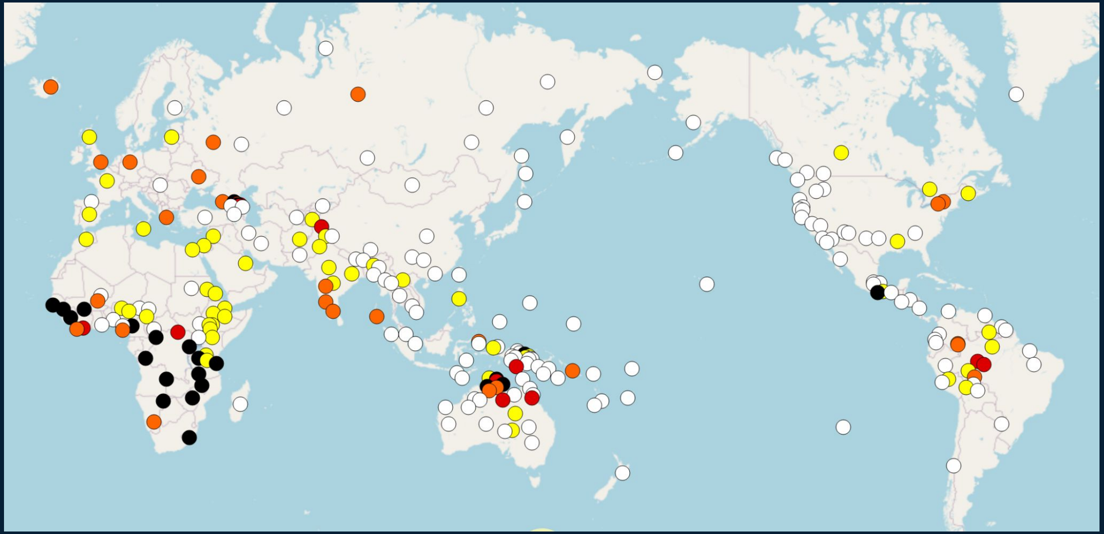


Image courtesy of *Yulia Tsvetkov*

Transfer of **resources** and **models** from **resource-rich source** to **resource-poor target languages**

- **Zero-shot learning**: train a model in one language/domain and assume it generalizes out-of-the-box in a low-resource language/domain
- **Few-shot learning**: train a model in one language/domain and use only few examples from a low-resource language/domain to adapt it

How do you define similar languages? Typology



Values

None	145	
Two	50	
Three	26	
Four	12	
Five or more	24	

reload

THE WORLD ATLAS OF LANGUAGE STRUCTURES ONLINE

NLLB Approach

1. **Acquire more data**
    - Collect human translations for **training** and **evaluation**
    - Innovate in large-scale data mining across the web

2. **Adapt massively multilingual systems**
    - Utilise cross-lingual transfer to allow related languages to learn from one another
    - Better than bilingual models, but enabling representation of hundreds of languages while retaining strong translation quality is difficult

NLLB Approach

- **Create professionally (human-)translated datasets**
    - Evaluation datasets for translation quality (FLORES-200, Toxicity-200, NLLB-MD)
    - Training datasets (NLLB-SEED)
- **Develop tools for large scale data mining**
- **NMT model developments**
    - NLLB-200: Sparsely Gated MoE model (with regularisation) for machine translation

Creating Professionally Translated Datasets

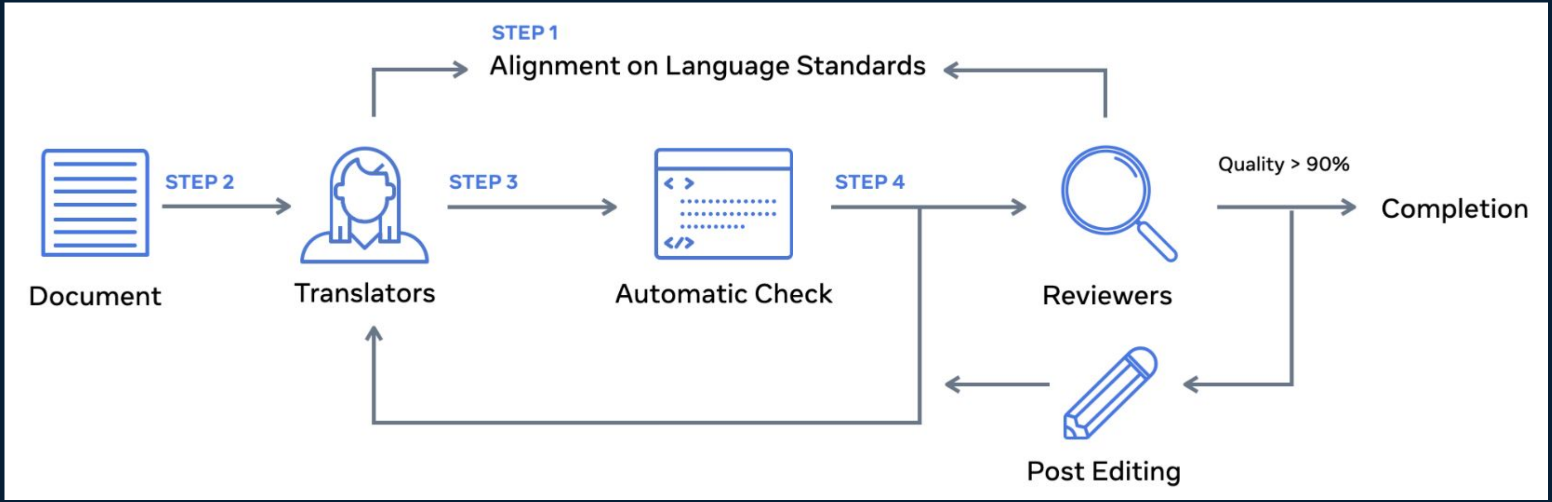
- **FLORES-200:**
  - Machine translation research requires the development of high-quality evaluation / benchmark datasets to assess progress
- **NLLB-SEED:**
  - Machine learning is notoriously data-hungry - for generation tasks like translation, require some high quality starter data
- **NLLB-MD:**
  - Avoiding overfitting and achieving strong out-of-domain performance is a major challenge in machine translation

FLORES-200

- Many-to-many translation benchmark to measure translation quality through **40602** translation directions

- **3001** sentences sampled from English-language Wikimedia projects (Wikinews, Wikijunior, Wikivoyage)

FLORES-200



STEP 1
Alignment on Language Standards
STEP 2
STEP 3
STEP 4
Document
Translators
Automatic Check
Reviewers
Quality > 90%
Completion
Post Editing

NLLB-SEED

- **T**raining set of professionally-translated sentences in Wikipedia domain in **39** languages
  - Consists of around 6000 sentences from English Wikipedia articles

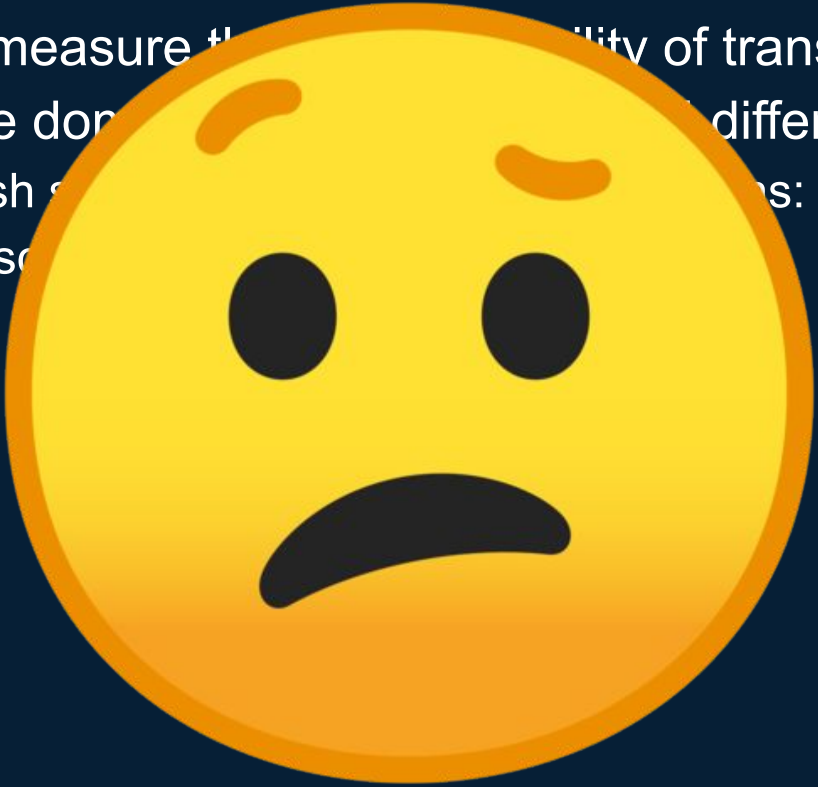
- Mostly used for training data rather than model evaluation
  - Did not go through the same quality assurance processes as FLORES-200 did

NLLB-MD

- Evaluation to measure the generalisability of translation models across multiple domains in **6** languages in **4** different domains
    - 3000 English sentences in the following domains: news, scripted formal speech, unscripted formal speech, health

NLLB-MD

- Evaluation to measure t… …ility of translation models across multiple do… …different domains
  - 3000 English … …s: news, scripted formal speech, unsc…



In sum..

- NLLB-SEED only contains translations in 39 languages, but NLLB-200 is a translation model for over 200 languages
  - How can we collect more data???

- Current techniques used for training translation models are difficult to extend to low-resource settings
  - Both aligned textual data (*bitext*: pairs of translated sentences) and single language data (*monolingual*) is limited

- Many low-resource languages are supported only through small targeted bitext dataset such as the Bible
  - Extremely limited in domain diversity

Automatically Creating Translation Data

1. Extend existing datasets by collecting non-aligned monolingual data
2. Use large-scale data mining to identify sentences that have high probability of being translations of each other in different languages

To do this, we need to:

1. Develop language identification (LID) systems to accurate label which language a given piece of text is written in
2. Gather and clean monolingual data at scale
3. Gather **bitexts** by using a sentence encoding approach to determine whether two sentences are parallel or not

1. Language Identification

- **Language identification (LID) challenges:**
  - Domain mismatch could occur due to the scarcity of text reliably labelled by language
  - Severe class imbalance as many low-resource languages of interest have low presence on the web
  - Efficiency of current approaches to run over large web collections is low even though they are massively parallelisable

- **NLLB Solutions:**
  - Use **fasttext** (tool Meta created for LID)
    - Widely used for text classification due to its speed while achieving good quality
  - Use FLORES-200 development set (⅓ of the dataset) to tune the model
  - In tuning, upsampled under-represented languages to combat massive class imbalance

2. Gathering monolingual data

Sentences are extremely noisy - often have URLs or hashtags which confuse the LID and script identification

- **Use web data from CommonCrawl and ParaCrawl**
  - Pre-process to remove markup and stripping HTML
  - Convert raw web text in paragraph form to sentences
    - Apply language identification to each web paragraph
    - Apply sentence splitting based on the language
- **Raw paragraphs may contain mix of languages or include code switching**
  - To avoid having mix of languages, re-run LID to identify the language of the sentence
  - If sentence-level LID not equal to paragraph-level LID, discard the sentence to ensure we only keep high-confidence sentences in the target language
  - Also discard if do not use the expected script for the target language
- **Apply some heuristics for data cleaning that don't match reasonable quality criteria (minimum/maximum length, space/punctuation/number/emoji ratios, maximum number of repeated characters, etc.)**
- **Run deduplication**
- **For high resource languages, run sentences through LM to see if they're reasonable**

2. Gathering monolingual data

- **Use web data from CommonCrawl and ParaCrawl**
  - Pre-process to remove markup and stripping HTML
  - C
- **Ra hing**
  -
  - only keep
  -
- **App quality criter oji ratios, maximum number of repeated characters, etc.)**
- **Run deduplication**
- **For high resource languages, run sentences through LM to see if they're reasonable**

- **Processed about 37.7 PB of data!**

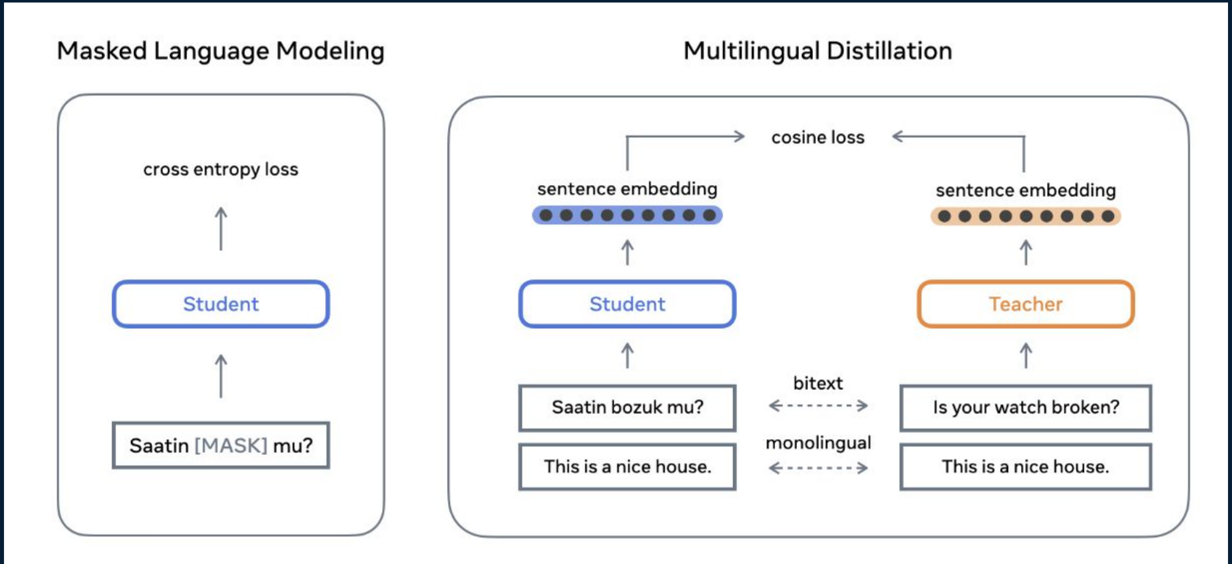
3. Bitext mining

- **Existing parallel corpora for low-resource languages often take from known collections of multilingual content, e.g. Bible or publications of multinational organisations**
  - Limited in quality and domain
- **NLLB automatically creates translation training datasets through bitext mining**
  - Focus on bitexts paired with English, but in the future interested in mining through other language pairs
- **Approach:**
  - Learn a multilingual embedding space
  - Use a similarity measure to decide whether two sentences are parallel or not
  - This can be applied to all possible pairs in two collections of monolingual texts
- **Challenges:**
  - How can you make sure all languages are well-learned?
  - How can we account for large imbalances in available training data?
  - Training a massively multilingual sentence encoder from scratch each time a new set of languages is added is computationally expensive and wasteful
    - Can we develop an approach to progressively add low-resource languages without needing to retrain the full model from scratch?

3. Bitext mining

- **Adopt a student-teacher mining approach:**
  - English-paired bitexts are used to both learn the English embedding space of the monolingual teacher while using the non-English side to learn a new language
  - Students are specialised for one language or several similar languages
  - Students are randomly initialised to handle low-resource languages which we don't have a pre-trained LM
  - Students may have dedicated tokenizer to accommodate scripts and tokens in student languages
  - Students learn by minimising the cosine loss with the teacher
  - Students can also have a masked language modelling loss to leverage student language monolingual data
- **All students trained on available bitexts in their languages complemented with two million sentences of English/English and English/Spanish**
  - Aim to "anchor" the students to the English embedding space and to make it more robust by including English/Spanish bitexts to jointly learn new languages
- **Resulting student encoders for 148 languages are called LASER3**

3. Bitext mining

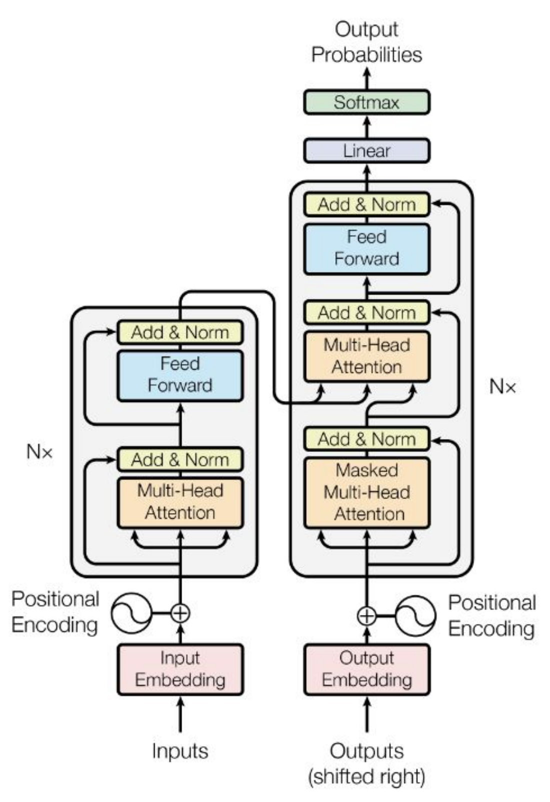


3. Bitext mining

- Once we have sentence encoders in multiple languages, mined 148 bitexts pairs with English to total 761 million sentence pairs
- Two sentences in different languages are considered pairs if they have high similarity in the sentence encoding space
- Limitations
  - Still limited by lack of monolingual data for low-resource languages
  - They can have low presence on the web and the data we curate has several filtering stages (LID, aggressive filtering/cleaning, differing domains) resulting in a lack of mined bitext pairs for a language
  - Still use all available bitext to train such as Bible

Modelling

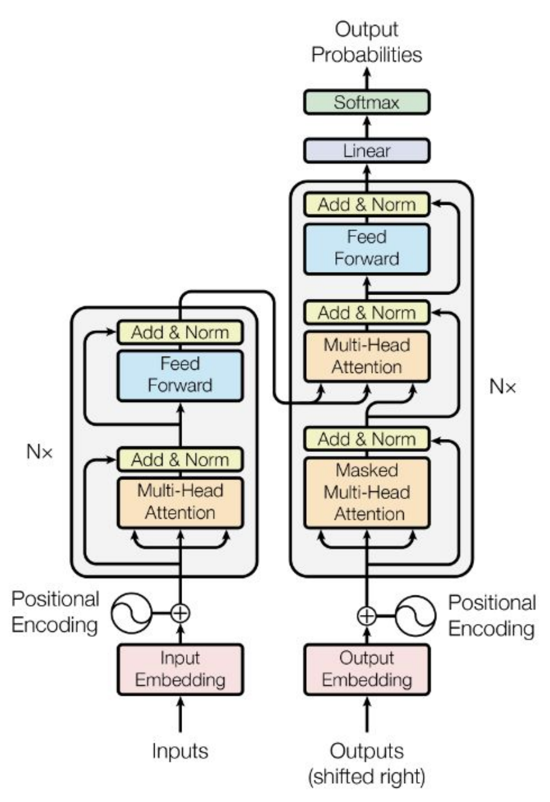
- Model multilingual neural machine translation as a **sequence-to-sequence** task
- Maximising the probability of the translation in target language **T** given the source sentence **S**, the source language **l_{s}** and target language **l_{t}**:
  - **P(T | S, l_{s}, l_{t})**
- Model architecture is (generally) based on **Transformer encoder-decoder** architecture



Transformer Encoder-Decoder

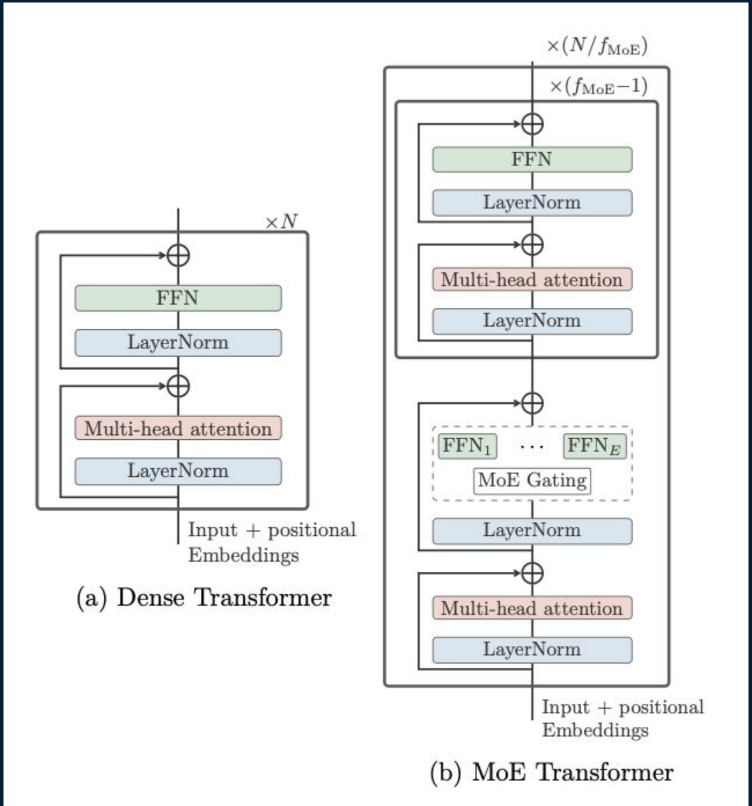
- **Encoder**: takes sequence of tokens **W** and source language **l_{s}** and outputs sequence of embeddings **H**
- **Decoder**: takes sequence of embeddings **H** and target language **l_{t}** to produce target tokens **V**

$$
\begin{aligned}
H &= \texttt{encoder}(W,\ \ell_s), \\
\forall i \in [1,\ldots,T],\ v_{i+1} &= \texttt{decoder}(H,\ \ell_t,\ v_1,\ \ldots,\ v_i).
\end{aligned}
$$



Modelling

- Aim: train a massively multilingual translation model on to **handle many translation directions at once**
  - Beneficial for cross-lingual transfer between related languages
- Problem: can result in **increased interference** between unrelated languages
- Approach: **Sparsely Gated Mixture of Experts (MoE)** models

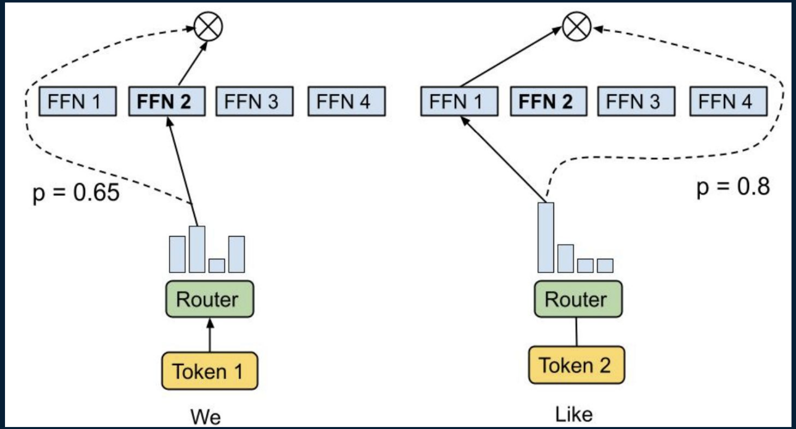


(a) Dense Transformer

(b) MoE Transformer

Mixture of Experts

- Type of **conditional compute model** that activate a **subset** of model parameters for a given input
- In contrast, **dense** models activate **all model parameters** per input
- For MoE transformer models, **replace** FFN layers with **multiple parallel expert** networks accompanied by a **gating network / router**



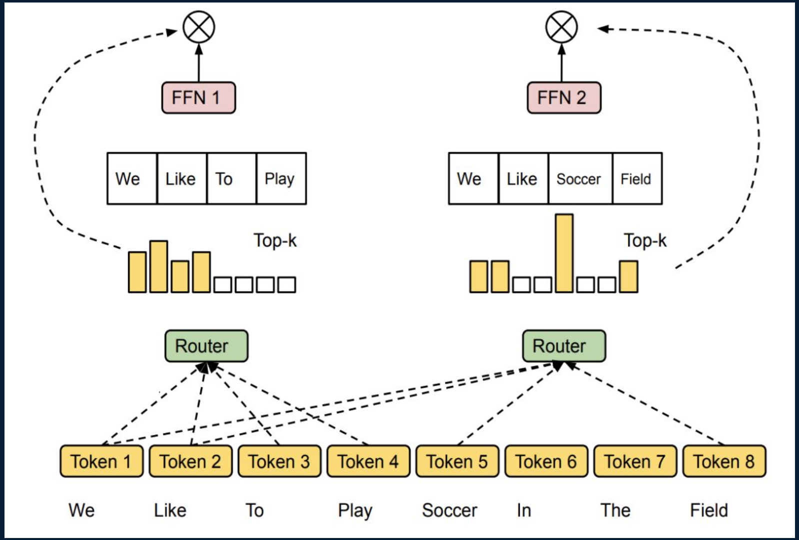
Sparsely Gated Mixture of Experts

- Significantly increases representational capacity while maintaining same inference and training efficiencies (in terms of FLOPs)
- Replace FFN sublayer with a MoE sublayer every f_{MOE} layers

$$\mathrm{FFN}_e(x_t) = W_o^{(e)}\,\mathrm{ReLU}(W_i^{(e)} \cdot x_t), \quad (\forall e \in \{1, \ldots, E\})$$

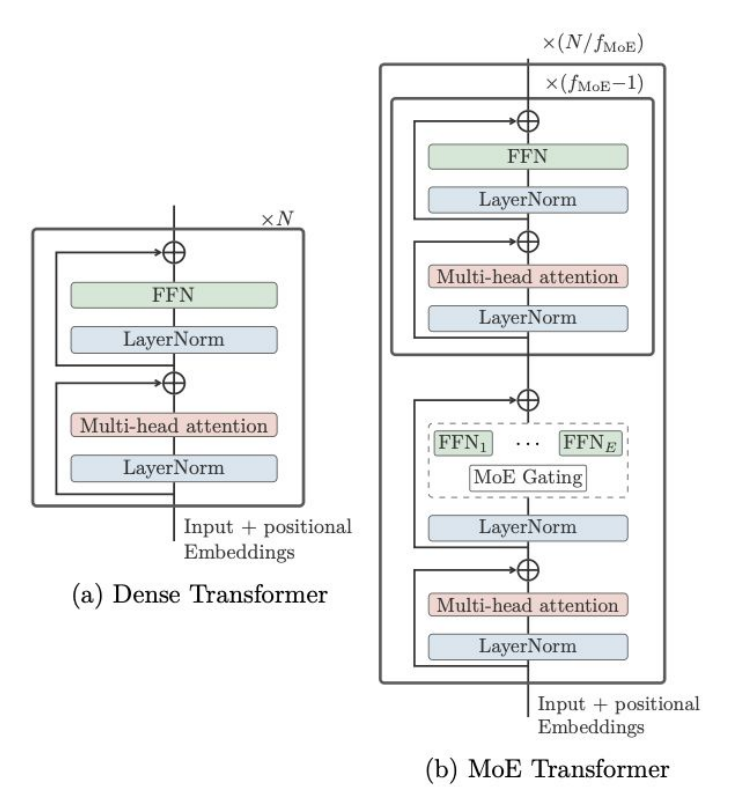
$$G_t = \mathrm{softmax}(W_g \cdot x_t), \quad \mathcal{G}_t = \text{Top-k-Gating}(G_t),$$

$$\mathrm{MoE}(x_t) = \sum_{e=1}^{E} \mathcal{G}_{te} \cdot \mathrm{FFN}_e(x_t),$$



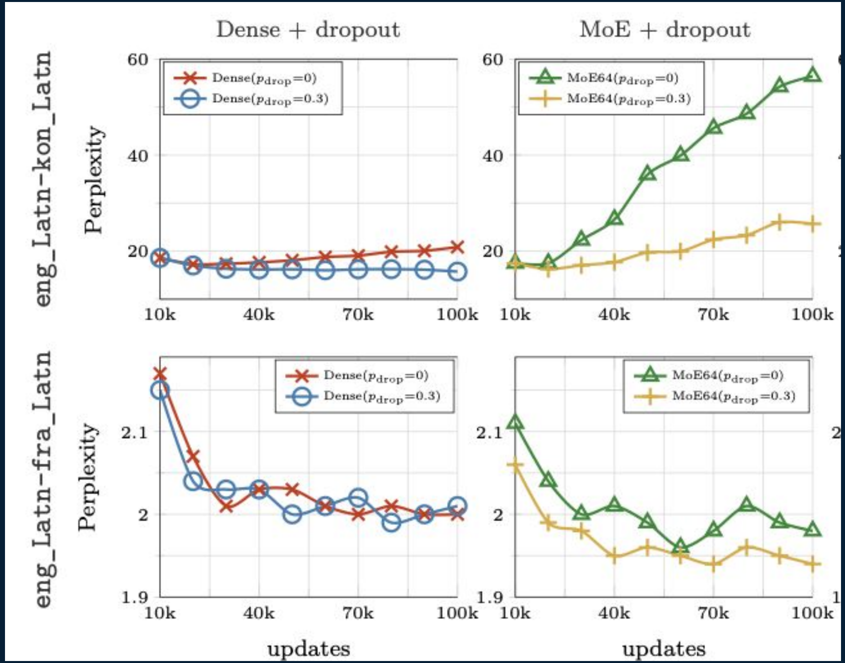
Sparsely Gated Mixture of Experts

- Each input token is routed to a number (**k=2**) “expert” sub-networks
    - Outputs are then **combined**
- Gating layer weights is simply another set of learnable weights
- Intuitively, individual experts learn from different inputs
- Motivation is to allow different parameters to model **different aspects** of the input space
    - Different parts of the model can focus on specialising in translating different languages



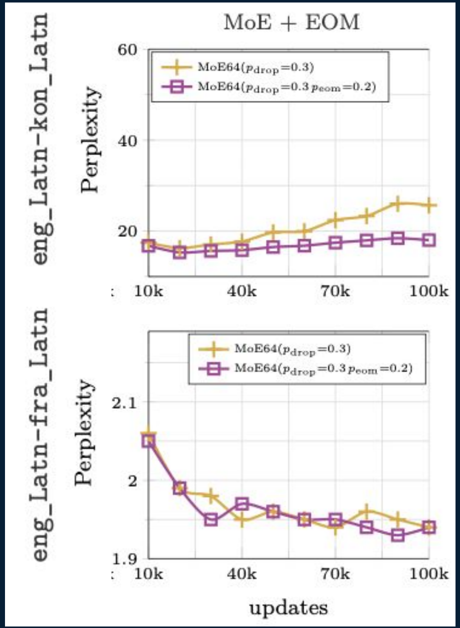
Some challenges with Mixture of Experts

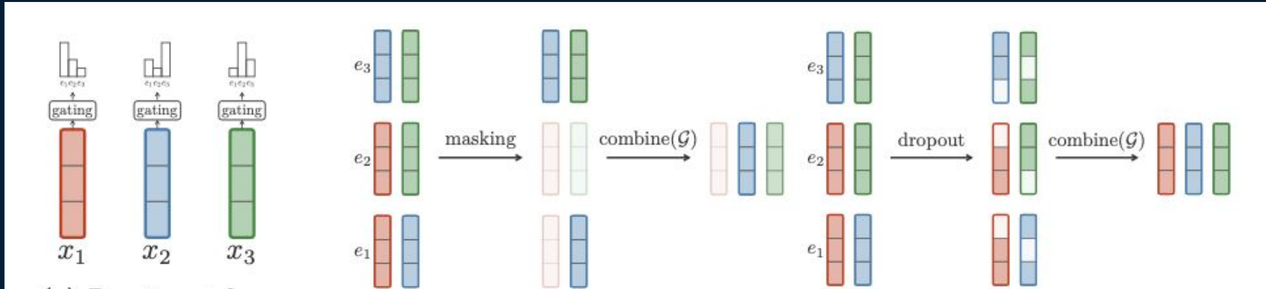
- **Load balancing**: how do we stop the gating network from routing all tokens to the same expert(s)?
  - **Expert capacity**: set a maximum number of tokens that a expert can process in a batch
  - **Auxiliary loss**: add a load-balancing loss to push tokens to be uniformly distuributed across experts
- Training instability and overfitting
  - MoE performance is significantly improved with increased **dropout**
  - Also used **Expert Output Masking (EoM)** and **Curriculum Learning** (introduce low-resource language pairs in later stages of model training)



Expert Output Masking

- **Regularisation** for massively multilingual MoE models
  - MoE models enable specialised expert capacity to be activated based on the input token - larger capacity can cause models to overfit (esp. on low-resource directions)
- MoE **Expert Output Masking** (EOM) is a technique where we **mask the expert output** for a random fraction, p_{EOM}, of input tokens
  - For token inputs with dropped experts, the first and/or second expert are effectively skipped

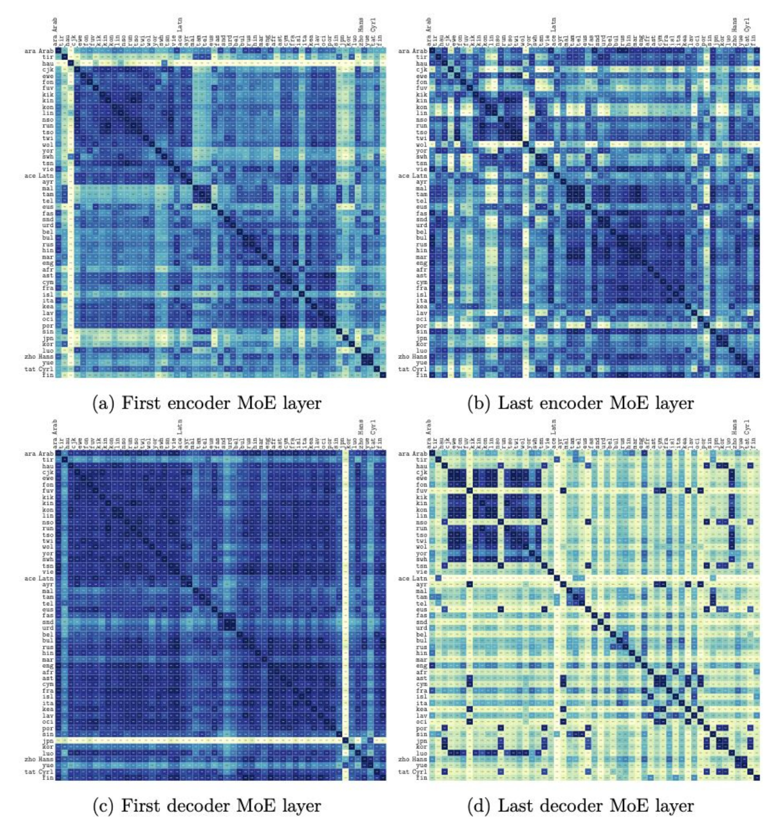




(a) Routing tokens (b) MoE Expert Output Masking (c) Overall dropout

What do multilingual Sparsely Gated MoEs learn?

- MoEs theoretically allow models to **specialise expert capacity** for different tasks / languages
- To assess what they learn:
  - Train and perform forward passes with FLORES-200 dev
  - For each task / language pair, log the routing decisions prior to Top-k gating to see how the model decides what experts to use per language
  - Look at cosine similarity between expert routing decisions for languages



(a) First encoder MoE layer

(b) Last encoder MoE layer

(c) First decoder MoE layer

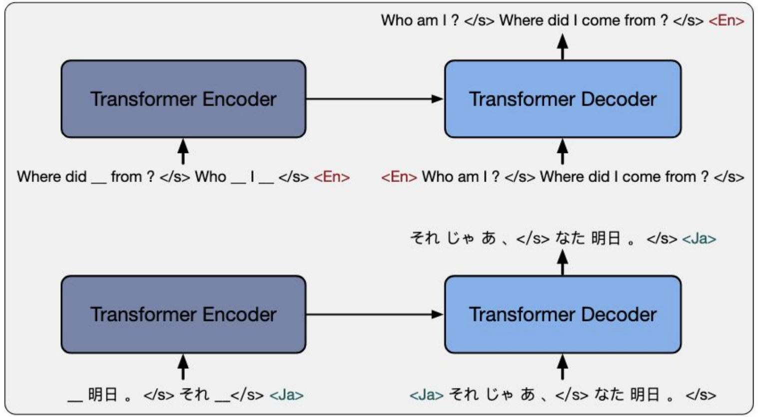
(d) Last decoder MoE layer

Data Augmentation

- For low-resource languages, there's generally limited or no bitext data available
  - If they are available, they could be in a very narrow domain or noisy
- Two ways to leverage monolingual data:
  - Incorporating **self-supervision learning (SSL)** tasks during training
  - **Backtranslation**: creating **synthetically generated** bitext / parallel corpora that are noisy on the source side via machine translation

Self-supervised learning

- Hope to learn patterns and constructs of a language from monolingual text
- Two approaches:
  - **Denoising Autoencoder (DAE)**:
    - Target: sentence from monolingual corpus
    - Input: **noised version** of the target monolingual sentence (randomly mask spans of text or replace with random tokens)
  - **Causal Language modelling (LM)**:
    - Simply train decoder on next token prediction (encoder / source is empty)



Multilingual Denoising Pre-Training (mBART)

	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

Backtranslation

- Another way to leverage monolingual data is through backtranslation:
  - Use an existing machine translation model to translate to obtain a noisy source pair
  - Create synthetically augmented data for translation models
- But for low-resource languages, MT models are often not good enough and generated data is noisy and degenerate

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

Backtranslation and Data Tagging

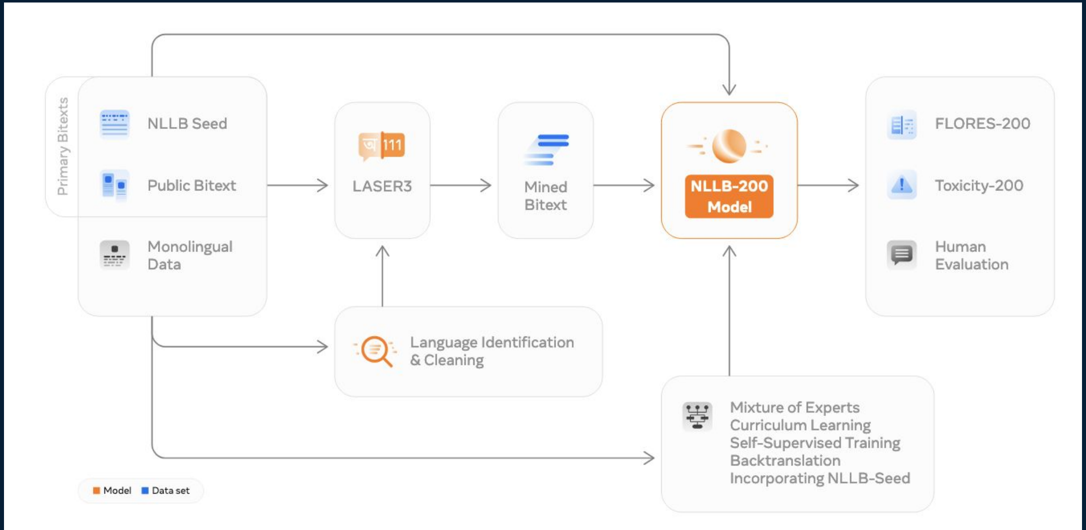
- Extra data from **backtranslation improves model performance** despite the data being noisy
  - Combination of different, complementary sources of noise is (potentially) why its addition is still beneficial to overall performance
- Additional performance is found when specifically **tagging** the data source
  - Introduce special tokens to tag the data, e.g. **<MINED_DATA>**, **<MMT_BT_DATA>**, **<SMT_BT_DATA>**
  - Useful for the model to distinguish between synthetic and natural data
  - Helps model from overfitting on it

	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**

	eng_Latn-xx				xx-eng_Latn				xx-yy
	all	high	low	v.low	all	high	low	v.low	all
No Tags	42.8	54.5	38.0	31.9	54.8	64.2	50.9	48.4	40.8
Single Tag	44.0	55.2	39.4	**34.2**	55.5	64.6	51.8	50.5	40.7
Finegrained Tags	**44.2**	**55.5**	**39.6**	34.0	**55.9**	**64.9**	**52.2**	**50.9**	**41.1**

Bringing it all together

- NLLB-200
    - Transformer encoder-decoder with **24 encoder** and **24 decoder** layers
    - Model dimension 2048
    - FFN dimension 8192
    - 16 Attention heads
    - Replace FFN with Sparsely Gated MoE layer **every 4th** Transformer block
    - **54.5B** parameters (but FLOPs similar to **3.3B** dense model)



This was a lot!

Questions?