Glot500: Scaling Multilingual Corpora and Language Models to 500 Languages (Area Chair Award)

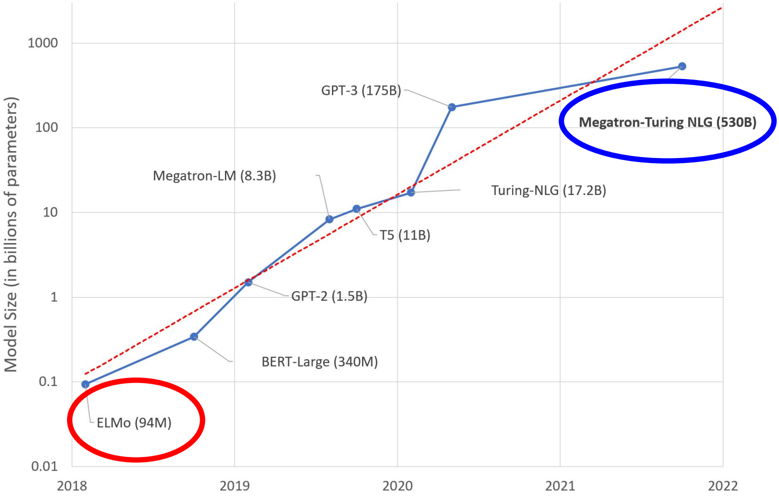
Ayyoob Imani*, Peiqin Lin*
Amir Hossein Kargaran, Silvia Severini, Masoud Jalili Sabet
Nora Kassner, Chunlan Ma, Helmut Schmid
André F. T. Martins, François Yvon, Hinrich Schütze

Jul 11, 2023

CIS
LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN
mcml Munich Center for Machine Learning
Unbabel
ISIR INSTITUT DES SYSTÈMES INTELLIGENTS ET DE ROBOTIQUE

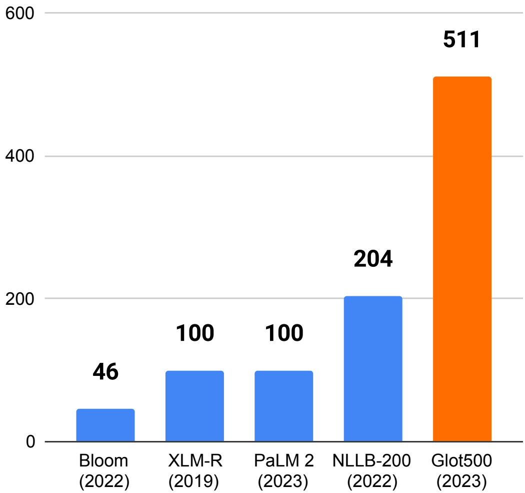
Scaling Large Language Models Vertically

Increasing Model Size from 2018-2022



Scaling Large Language Models Horizontally

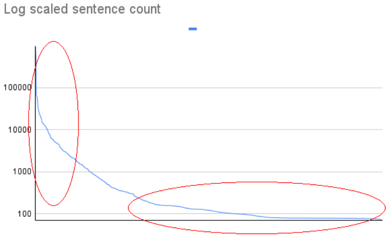
Public multilingual language models



Long-tail distribution

Among 7000 languages:

- head languages (100)
  - Covered by XLM-R
  - Large corpora available
- tail languages (1000s)
  - not covered by XLM-R
  - Little data available



Glot500

- Data: a corpus covering 2000+ languages → Glot2000-c
- Model: an LLM covering 511 languages → Glot500-m
- Evaluation: Evaluate Glot500-m on a diverse suite of tasks

Glot500: Data Collection

A lightweight method: Benefit from previous efforts

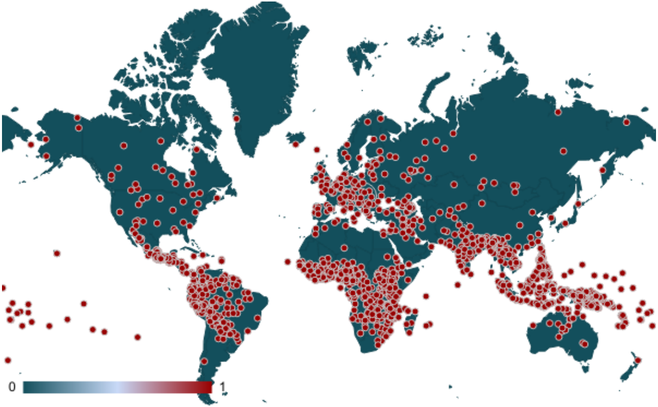
- **Websites**, e.g., Jw.org, lyricstranslate.com
- **Datasets (150 datasets)**
  - **Multilingual**, e.g., mC4, Oscar, MTData, Tatoeba
  - **Single language or single family**, e.g., Indic NLP

No Language ID required!

Glot500: Data Cleaning

- Sentence level filters
  - Character repetition
  - Word repetition
  - Special characters
  - Small sentences
  - Duplicates
- Corpus level filters
  - Language script mismatch
  - Perplexity mismatch

Glot500: Data



- 2266 languages
- Worldwide

Glot500: Model

Training Data

- $>$30k sentences
- 511 languages (both head and tail languages)
- 534 language-scripts
- 610 GB

Glot500: Model

Training Model

- XLM-R Base as the starting point
- Vocabulary extension: 250K + 150K (new) = 400K
- Continued Pretraining: Masked Language Modeling

Glot500: Model Size Comparison

	XLM-R-B	XLM-R-L	Glot500-m
Model Size	278M	560M	395M
Vocab Size	250K	250K	401K
Transformer Size	86M	303M	86M

- Glot500-m and XLM-R-B have the **same transformer size**
- Glot500-m has a **larger vocabulary**, resulting in an overall **larger model**
- Glot500-m is smaller than XLM-R-L

Glot500: Downstream Tasks

	head	tail	measure (%)
Sentence Retrieval Tatoeba	70	28	Top10 Acc.
Sentence Retrieval Bible	94	275	Top10 Acc.
Text Classification (Taxi1500)	90	264	F1
NER	89	75	F1
POS	63	28	F1
Roundtrip Alignment	85	288	Accuracy

- **427 (80%)** language-scripts evaluated by at least one task
- More than any prior work

Glot500: Main Results on Tail Languages

	tail		
	XLM-R-B	XLM-R-L	Glot500-m
Pseudoperplexity	304.2	168.6	**12.2**
Sentence Retrieval Tatoeba	32.6	33.6	**59.8**
Sentence Retrieval Bible	7.4	7.1	**43.2**
Text Classification	13.7	13.9	**46.6**
NER	47.5	51.8	**60.7**
POS	41.7	43.5	**62.3**
Roundtrip Alignment	2.6	3.1	**4.5**

For tail languages

- Glot500-m > XLM-R-B in all tasks
- Glot500-m > XLM-R-L in all tasks

Glot500: Main Results on Head Languages

	head		
	XLM-R-B	XLM-R-L	Glot500-m
Pseudoperplexity	12.5	**8.4**	11.8
Sentence Retrieval Tatoeba	66.2	71.1	**75.0**
Sentence Retrieval Bible	54.2	58.3	**59.0**
Text Classification	51.3	**60.5**	54.7
NER	61.8	**66.0**	63.9
POS	76.4	**78.4**	76.0
Roundtrip Alignment	3.4	4.1	**5.5**

For head languages

- Glot500-m $>$ XLM-R-B in all tasks except POS
- Glot500-m $>$ XLM-R-L in 3/7 tasks

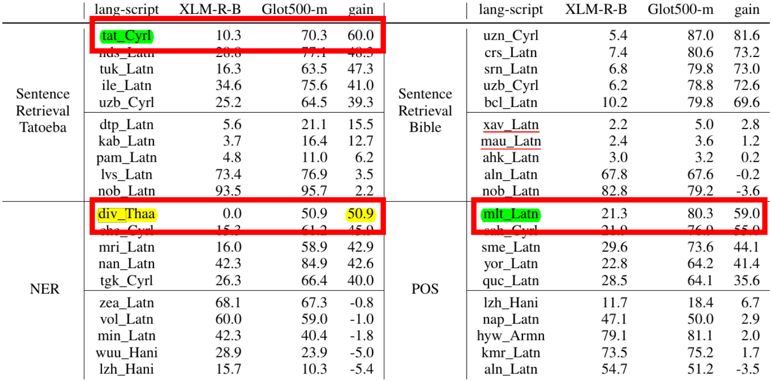
Glot500: Main Results on All Languages

	all		
	XLM-R-B	XLM-R-L	Glot500-m
Pseudoperplexity	247.8	136.4	**11.6**
Sentence Retrieval Tatoeba	56.6	60.4	**70.7**
Sentence Retrieval Bible	19.3	20.1	**47.3**
Text Classification	23.3	25.8	**48.7**
NER	55.3	59.5	**62.4**
POS	65.8	67.7	**71.8**
Roundtrip Alignment	2.8	3.3	**4.7**

For all languages

- Glot500-m $>$ XLM-R-B in all tasks
- Glot500-m $>$ XLM-R-L in all tasks

Glot500: Languages with Big Gains

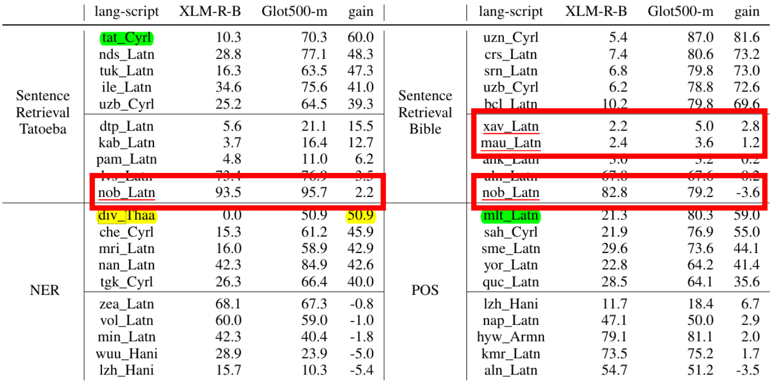


	lang-script	XLM-R-B	Glot500-m	gain		lang-script	XLM-R-B	Glot500-m	gain
Sentence Retrieval Tatoeba	tat_Cyrl	10.3	70.3	60.0	Sentence Retrieval Bible	uzn_Cyrl	5.4	87.0	81.6
	nds_Latn	28.8	77.1	48.3		crs_Latn	7.4	80.6	73.2
	tuk_Latn	16.3	63.5	47.3		srn_Latn	6.8	79.8	73.0
	ile_Latn	34.6	75.6	41.0		uzb_Cyrl	6.2	78.8	72.6
	uzb_Cyrl	25.2	64.5	39.3		bcl_Latn	10.2	79.8	69.6
	dtp_Latn	5.6	21.1	15.5		xav_Latn	2.2	5.0	2.8
	kab_Latn	3.7	16.4	12.7		mau_Latn	2.4	3.6	1.2
	pam_Latn	4.8	11.0	6.2		ahk_Latn	3.0	3.2	0.2
	lvs_Latn	73.4	76.9	3.5		aln_Latn	67.8	67.6	-0.2
	nob_Latn	93.5	95.7	2.2		nob_Latn	82.8	79.2	-3.6
NER	div_Thaa	0.0	50.9	50.9	POS	mlt_Latn	21.3	80.3	59.0
	che_Cyrl	15.3	61.2	45.9		sah_Cyrl	21.9	76.9	55.0
	mri_Latn	16.0	58.9	42.9		sme_Latn	29.6	73.6	44.1
	nan_Latn	42.3	84.9	42.6		yor_Latn	22.8	64.2	41.4
	tgk_Cyrl	26.3	66.4	40.0		quc_Latn	28.5	64.1	35.6
	zea_Latn	68.1	67.3	-0.8		lzh_Hani	11.7	18.4	6.7
	vol_Latn	60.0	59.0	-1.0		nap_Latn	47.1	50.0	2.9
	min_Latn	42.3	40.4	-1.8		hyw_Armn	79.1	81.1	2.0
	wuu_Hani	28.9	23.9	-5.0		kmr_Latn	73.5	75.2	1.7
	lzh_Hani	15.7	10.3	-5.4		aln_Latn	54.7	51.2	-3.5

Big gains

- **New script**: Dhivehi (div_Thaa)
- **Big corpus size**: Tatar (tat_Cyrl), Maltese (mlt_Latn)

Glot500: Languages with No Gain



	lang-script	XLM-R-B	Glot500-m	gain		lang-script	XLM-R-B	Glot500-m	gain
Sentence Retrieval Tatoeba	tat_Cyrl	10.3	70.3	60.0	Sentence Retrieval Bible	uzn_Cyrl	5.4	87.0	81.6
	nds_Latn	28.8	77.1	48.3		crs_Latn	7.4	80.6	73.2
	tuk_Latn	16.3	63.5	47.3		srn_Latn	6.8	79.8	73.0
	ile_Latn	34.6	75.6	41.0		uzb_Cyrl	6.2	78.8	72.6
	uzb_Cyrl	25.2	64.5	39.3		bcl_Latn	10.2	79.8	69.6
	dtp_Latn	5.6	21.1	15.5		xav_Latn	2.2	5.0	2.8
	kab_Latn	3.7	16.4	12.7		mau_Latn	2.4	3.6	1.2
	pam_Latn	4.8	11.0	6.2		ahk_Latn	3.0	3.2	0.2
	lvs_Latn	73.4	76.9	3.5		aln_Latn	67.8	67.6	-0.2
	nob_Latn	93.5	95.7	2.2		nob_Latn	82.8	79.2	-3.6
NER	div_Thaa	0.0	50.9	50.9	POS	mlt_Latn	21.3	80.3	59.0
	che_Cyrl	15.3	61.2	45.9		sah_Cyrl	21.9	76.9	55.0
	mri_Latn	16.0	58.9	42.9		sme_Latn	29.6	73.6	44.1
	nan_Latn	42.3	84.9	42.6		yor_Latn	22.8	64.2	41.4
	tgk_Cyrl	26.3	66.4	40.0		quc_Latn	28.5	64.1	35.6
	zea_Latn	68.1	67.3	-0.8		lzh_Hani	11.7	18.4	6.7
	vol_Latn	60.0	59.0	-1.0		nap_Latn	47.1	50.0	2.9
	min_Latn	42.3	40.4	-1.8		hyw_Armn	79.1	81.1	2.0
	wuu_Hani	28.9	23.9	-5.0		kmr_Latn	73.5	75.2	1.7
	lzh_Hani	15.7	10.3	-5.4		aln_Latn	54.7	51.2	-3.5

No gain

- **Similar head language**: Norwegian Bokmål (nob_Latn)
- **Very small corpus**: Xavánte(xav_Latn)
- **Isolated language**: Huautla Mazatec (mau_Latn)

Glot500: Curse/Blessing of Multilinguality

lang-script	**Glot+1**	**Glot500-m**
Curse of Multilinguality		
rug_Latn, Roviana	**51.0**	49.0
yan_Latn, Mayangna/Sumo	**46.4**	31.8
wbm_Latn, Wa/Va	**49.6**	46.4
Blessing of Multilinguality		
ctd_Latn, Tedim Chin	47.4	**59.4**
quh_Latn, Southern Quechua	33.4	**56.2**
tat_Cyrl, Tatar	58.8	**67.2**

Glot+1 (Adapt to 1 lang) vs Glot500-m (Adapt to 500+ langs)

- Isolate languages → Curse of Multilinguality
- Support Through Related Languages → Blessing of Multilinguality

Glot500

Github (Code, Data, Model)

https://github.com/cisnlp/Glot500

See you again on Jul 12 (Wed) 9am at Bay - Unit 3!