ISIR
INSTITUT
DES SYSTÈMES
INTELLIGENTS
ET DE ROBOTIQUE

Building and Evaluating Massively Multilingual Language Models

François Yvon - CNRS

Nancy, journées du GDR LIFT, november 19th, 2023

SORBONNE UNIVERSITÉ

cnrs

Inserm
La science pour la santé
From science to health

Large Language Models

What we talk about when we talk about LLMs

(1) **LLMs model text** and can be used to **generate it** based on input context, i.e. by selecting the tokens that are the most likely given the partial context provided as input (either masked or as a prompt). The text can be in any modality — as characters, audio waves, pixels, etc.

(2) **LLMs receive large-scale pretraining**, where 'large-scale' refers to the pre-training data rather than the number of parameters. The exact threshold for LLM-qualifying volume of data is necessarily arbitrary, and we propose setting it to 1B tokens (inspired by Chelba et al. (2013)). As such, we consider BERT (Devlin et al. 2019) to be an LLM, given that it was trained on a corpus containing 3.3B tokens.

(3) LLMs **make inferences based on transfer learning**: LLMs are meant to be adaptable to many tasks, given that their language modeling objective (i.e. pre-training) encodes information that can then be leveraged in other tasks. This adaptation can be done in different ways – currently, two of the most popular ones are fine-tuning, as in BERT (Devlin et al. 2019), or prompting, as in GPT-3 (Brown et al. 2020a).

Mind your Language (Model): Fact-Checking LLMs and their Role in NLP Research and Practice (Luccioni, Rogers, 2023) arXiv e-prints, arXiv: 2308.07120

Monolingual LLMs

Learning LLMs parameters on large monolingual corpora with auxiliary tasks and natural annotations

1. Predict next word given prefix: pure decoder

Longtemps je me suis couché [mask] - unmask='de', train $P_\theta(w_t \mid w_{<t})$
(eg. GPT*, OPT, GPTJ, PALM*, LLAMA*)

2. Predict missing word given bidirectional contexts : pure encoder

Longtemps je me suis couché [mask] bonne heure- unmask='de', train $P_\theta(w_t \mid w_{\_t})$
(eg. BERT, Roberta, CamemBERT, FlauBERT, etc)

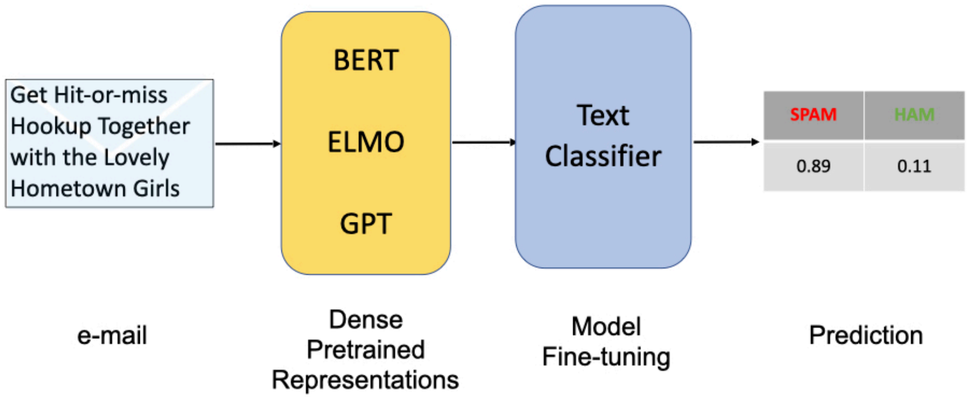
3. Denoising sequence to sequence : encoder-decoder

Longtemts je couché suis de bnone heur || Longtemps je me suis couché de bonne heure
train $P_\theta(w \mid \tilde{w}) = \prod_t P(w_t \mid w_{<t}, \tilde{w})$

(eg. BART, T5, etc)

Using Monolingual LLMs

1. **Fine-tune task-adapted model:** $h_{\phi,\theta} = h_\phi(f_\theta(w); c)$



2. **Multi-purpose text generation via prompting**

GEN	Of course. In Chorukor, Monday is ilopagar, Tuesday ilopager, . . .
	Wednesday ilopagur, Thursday ilopagir ...
Q&A	Answer this : What are the birth date and place of Ludvík Vaculík ? . . .
	23 July 1926, in Brumov, Moravia
SA	"This Czech writer has written some the most wonderful French novels."
	is a positive comment ? . . . [Yes | No]

The race for scores

81 models

AI21 Labs / J1-Jumbo v1 (178B)
AI21 Labs / J1-Large v1 (7.5B)
AI21 Labs / J1-Grande v1 (17B)
AI21 Labs / J1-Grande v2 beta (17B)
AI21 Labs / Jurassic-2 Jumbo (178B)
AI21 Labs / Jurassic-2 Grande (17B)
AI21 Labs / Jurassic-2 Large (7.5B)
Aleph Alpha / Luminous Base (13B)
Aleph Alpha / Luminous Extended (30B)
Aleph Alpha / Luminous Supreme (70B)
neurips / Local service
Anthropic / Anthropic-LM v4-s3 (52B)
Anthropic / Anthropic Claude 2.0
Anthropic / Anthropic Claude v1.3
Anthropic / Anthropic Claude Instant V1
UC Berkeley / Koala (13B)
BigScience / BLOOM (176B)
BigScience / BLOOMZ (176B)
BigScience / T0pp (11B)
BigCode / SantaCoder (1.1B)
BigCode / StarCoder (15.5B)
Cerebras / Cerebras GPT (6.7B)
Cerebras / Cerebras GPT (13B)
Cohere / Cohere xlarge v20220609 (52.4B)
Cohere / Cohere large v20220720 (13.1B)
Cohere / Cohere medium v20220720 (6.1B)
Cohere / Cohere small v20220720 (410M)
Cohere / Cohere xlarge v20221108 (52.4B)
Cohere / Cohere medium v20221108 (6.1B)
Cohere / Cohere Command beta (6.1B)
Cohere / Cohere Command beta (52.4B)

73 scenarios

Question answering
- MMLU
- BoolQ
- NarrativeQA
- NaturalQuestions (closed-book)
- NaturalQuestions (open-book)
- QuAC
- HellaSwag
- OpenbookQA
- TruthfulQA

Information retrieval
- MS MARCO (regular)
- MS MARCO (TREC)

Summarization
- CNN/DailyMail
- XSUM

Sentiment analysis
- IMDB

Toxicity detection
- CivilComments

Text classification
- RAFT

Aspirational scenarios
- Data-to-text generation
- Fact verification
- Copywriting
- Story generation

65 metrics

Accuracy
- none
- Quasi-exact match
- F1
- Exact match
- RR@10
- NDCG@10
- ROUGE-2
- Bits/byte
- Exact match (up to specified indicator)
- Absolute difference
- F1 (set match)
- Equivalent
- Equivalent (chain of thought)
- pass@1

Calibration
- Max prob
- 1-bin expected calibration error
- 10-bin expected calibration error
- Selective coverage-accuracy area
- Accuracy at 10% coverage
- 1-bin expected calibration error (after Platt scaling)
- 10-bin Expected Calibration Error (after Platt scaling)
- Platt Scaling Coefficient
- Platt Scaling Intercept

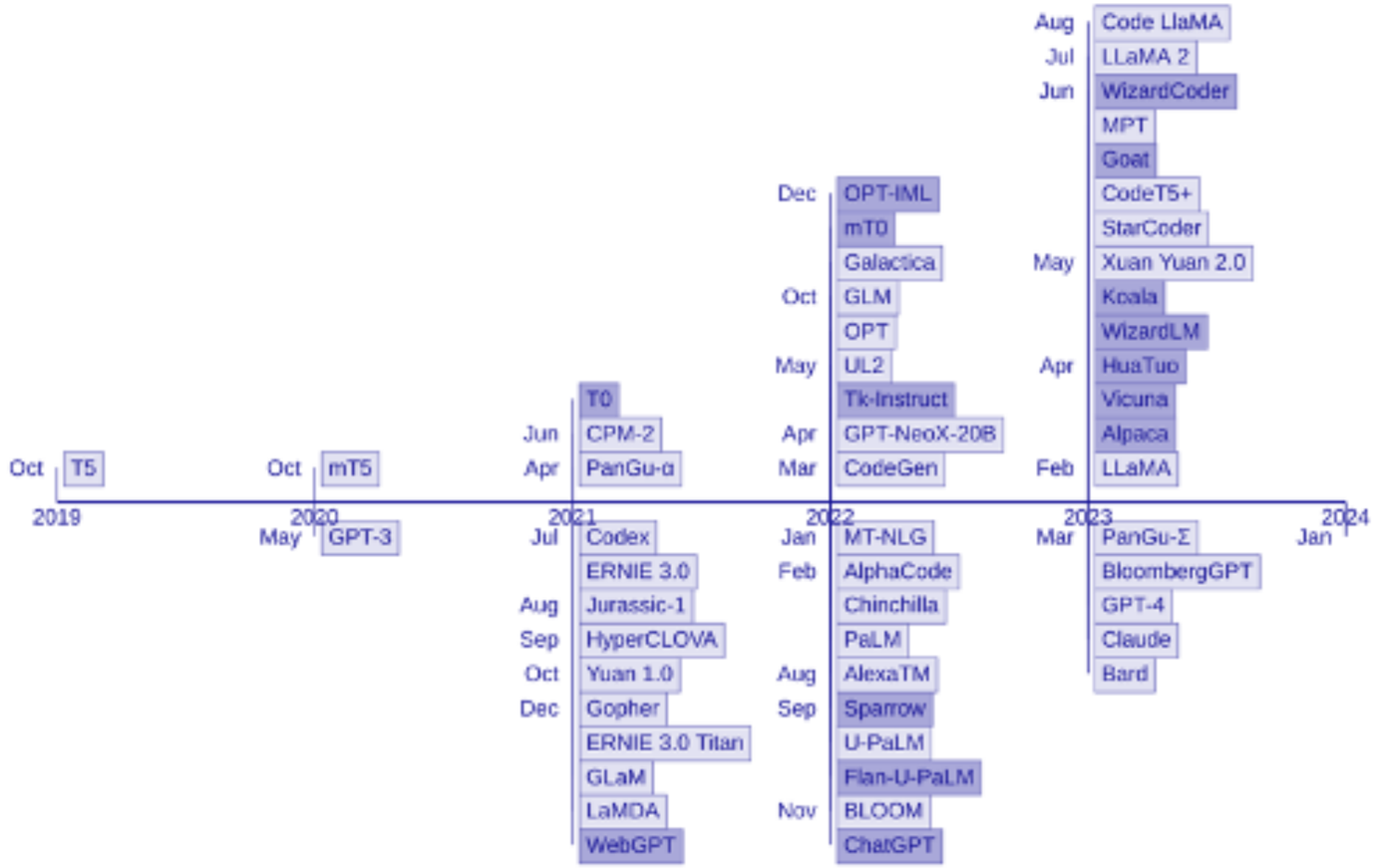
Robustness
- Quasi-exact match (perturbation: typos)
- F1 (perturbation: typos)
- Exact match (perturbation: typos)
- RR@10 (perturbation: typos)

Evaluation better be « holistic »

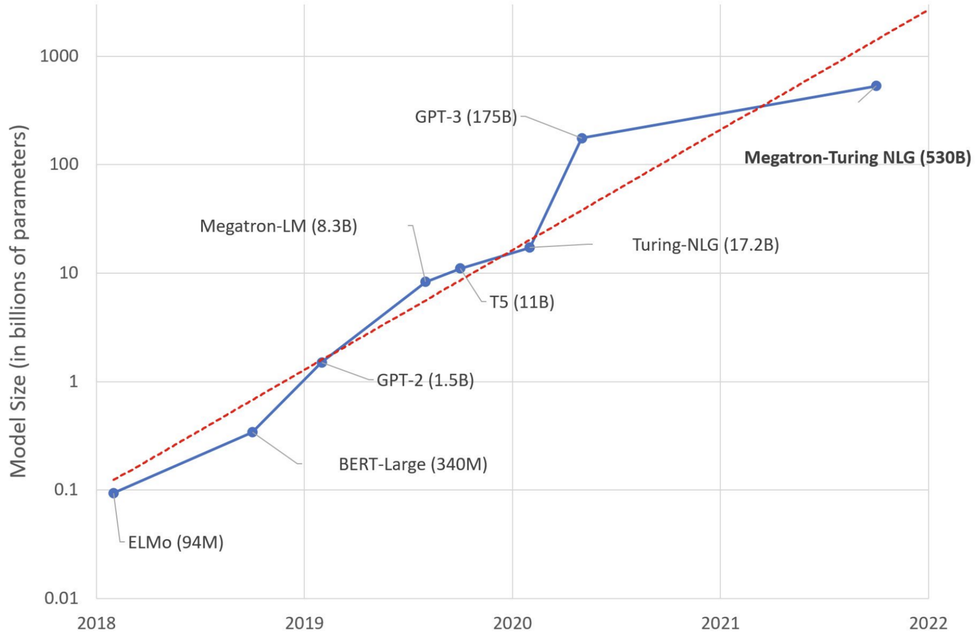
- tasks, bias, fairness, openness, etc.

Holistic Evaluation of Large Language Models (Liang et al, 2022)

The race for size



Naveed, H., Khan, A. U., Qiu, S., Saqib, M., Anwar, S., Usman, M., ... & Mian, A. (2023). A comprehensive overview of large language models. *arXiv preprint arXiv:2307.06435*.



Using deepspeed and Megatron to train Turing-NLG-530

Size related to « depth » and « thickness »

- Larger LLMs have better (emerging ?) « performance »
- Also: higher environmental footprint

Multilingual LLMs: mLLMs

Learning mLLMs parameters with large multilingual corpora, auxiliary tasks and natural annotations

1. Predict next word given prefix: pure decoder

(eg. mGPT, XGLM, BLOOM, PALM-2, Falcon)

2. Predict missing word given bidirectional contexts : pure encoder

(eg. mBERT, XLM-Roberta, etc)

3. Denoising sequence to sequence : encoder-decoder

(eg. mBART, mT5, etc)

Complementary objectives to bridge between languages

- parallel corpora (TLM loss, MT loss)
- bilingual dictionaries
- synthetic mixed-language data

mLLMs need multilingual texts

_Tous _les _être s _humain s _na issent _libre s _et _ég aux _en _digni té _et _en _droits . _Ils _sont _do u és _de _raison _et _de _conscience _et _doivent _agir _les _uns _en vers _les _autres _dans _un _ esprit _de _frater n ité .

_Všichni _lidé _rod í _se _svobod ní _a _sobě _rov ní _co _do _d ů stoj nosti _a _práv . _Jsou _na dán i _rozum em _a _s vědomí m _a _mají _spolu _jedna t _v _du chu _brat r ství .

_Tutti _gli _esse ri _umani _na scono _liberi _ed _e gu ali _in _digni tà _e _diritti . _Es si _sono _do tati _di _ragione _e _di _coscienza _e _devono _agir e _gli _uni _verso _gli _altri _in _spirito _di _fra tella nza .

Subword tokenizers are trainable modules

- require mixed-language, mixed-script training corpora
- larger language get more units, are better segmented, perform better
- parameter sharing for same-language scripts

Using Multilingual LLMs

1. Fine-tune task-adapted models on L1, use it to process L2 with zero-shot model transfer
Only requires annotation in L1

2. Multi-purpose, multilingual text generation via prompting

Translate into English

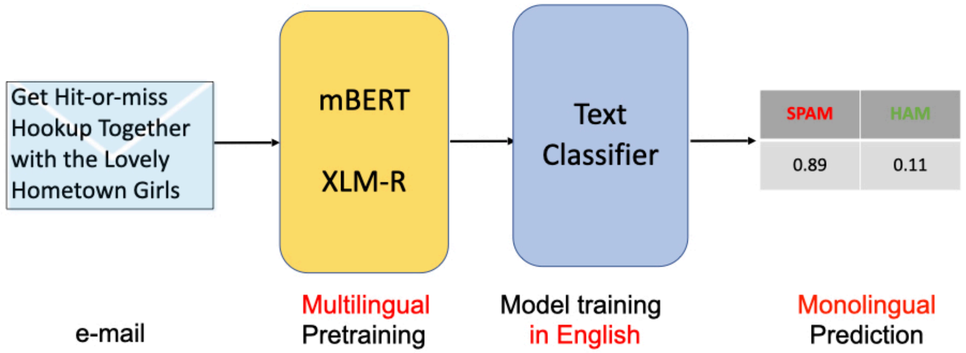
“ By the end of the year, we will have seven new pharmacists. ” :

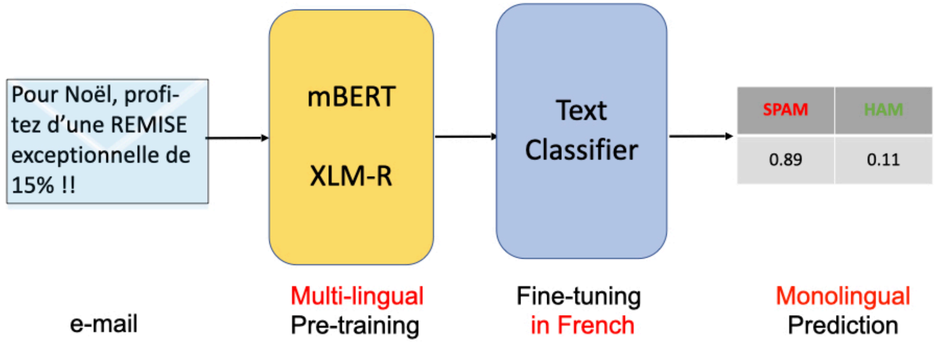
D’ici la fin de l’année, nous aurons sept nouveaux pharmaciens.

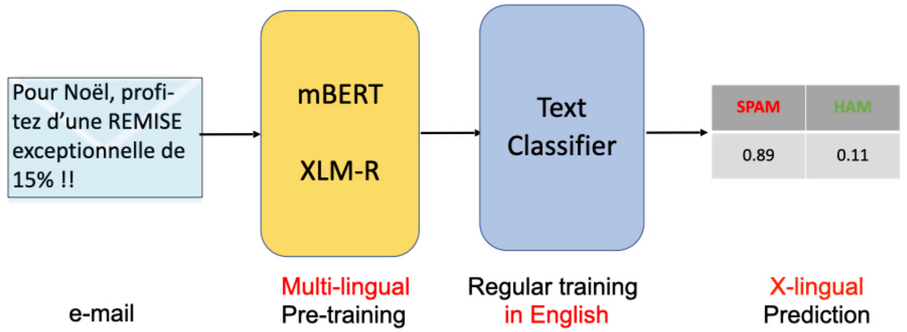
mLLMs are a blessing

- hardly more difficult than mLLMs
- excel in multilingual tasks
- enable X-lingual transfer

mReps are very useful

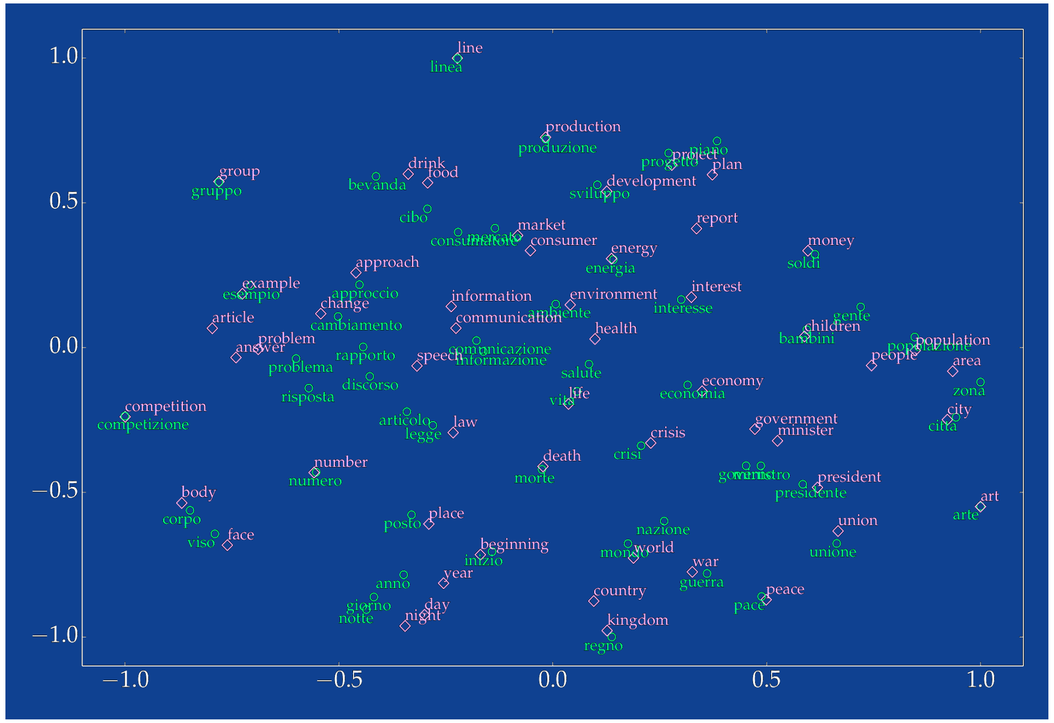






X-lingual transfer (zero-shot)

mReps are very useful



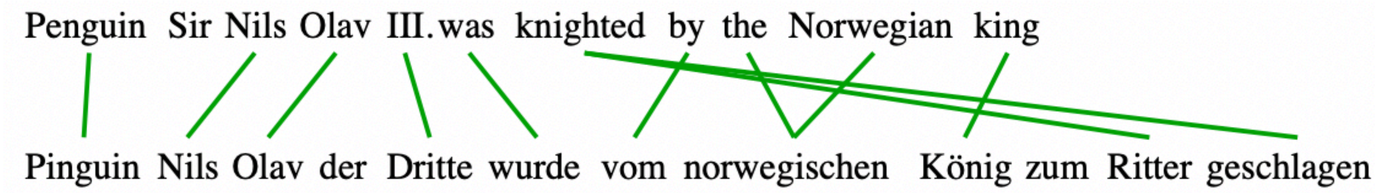
Unsupervised Cross-Lingual Representation Learning (Ruder et al., ACL 2019)

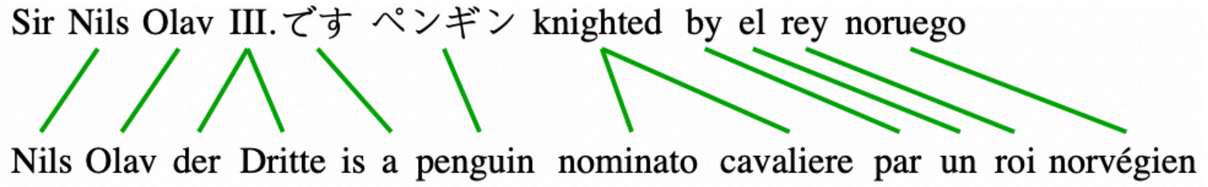
Multilingual representations make good alignments

Sentence level

In the gayest and happiest spirits she set forward with her father;	Elle partit avec son père, le visage souriant;
not always listening, but always agreeing to what he said;	elle n' écoutait pas toujours, mais elle acquiesçait de confiance.
They arrived .	Ils arrivèrent .
It is Frank and Miss Fairfax, said Mrs. Weston .	– C'est Frank et Mlle Fairfax, dit aussitôt Mme Weston .
I was just going to tell you of our agreeable surprize in seeing him arrive this morning.	– J'allai justement vous faire part de l'agréable surprise que nous avons eue en le voyant arriver.
He stays till tomorrow, and Miss Fairfax has been persuaded to spend the day with us .	Il reste jusqu'à demain et Mlle Fairfax a bien voulu, sur notre demande , venir passer la journée.

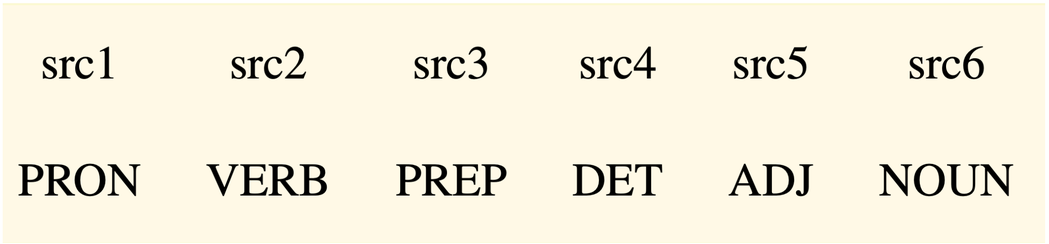
Word level





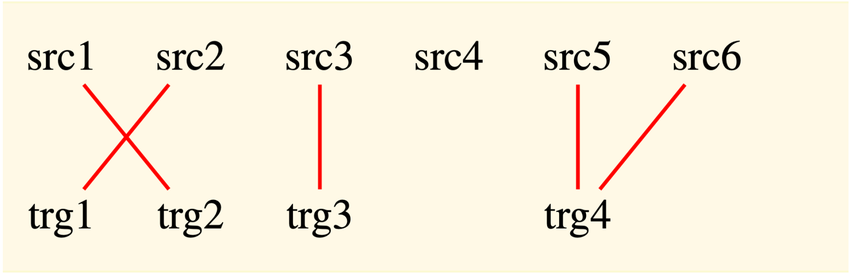
SimAlign: High Quality Word Alignments Without Parallel Training Data Using Static and Contextualized Embeddings (Jalili Sabet et al., Findings 2020)

mReps are very useful



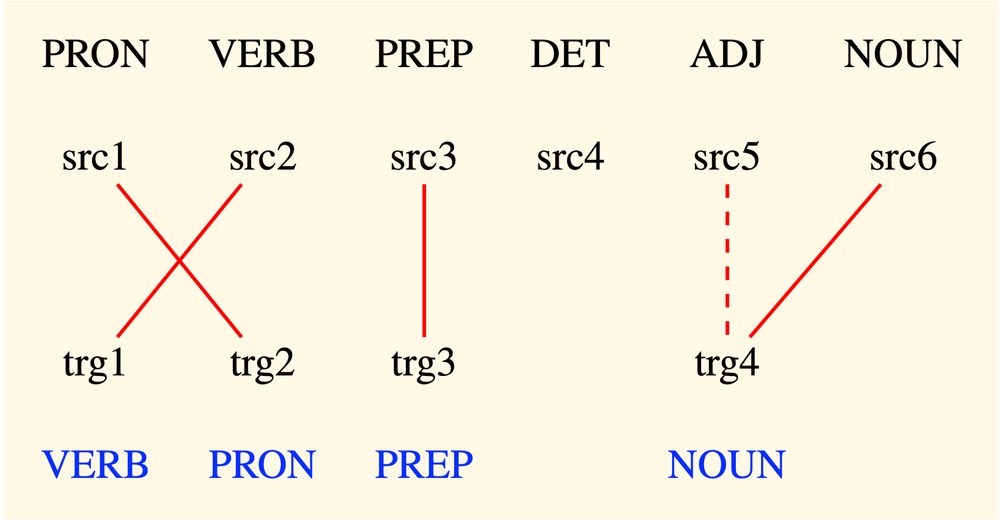
Tag source

Annotations for free via alignment links

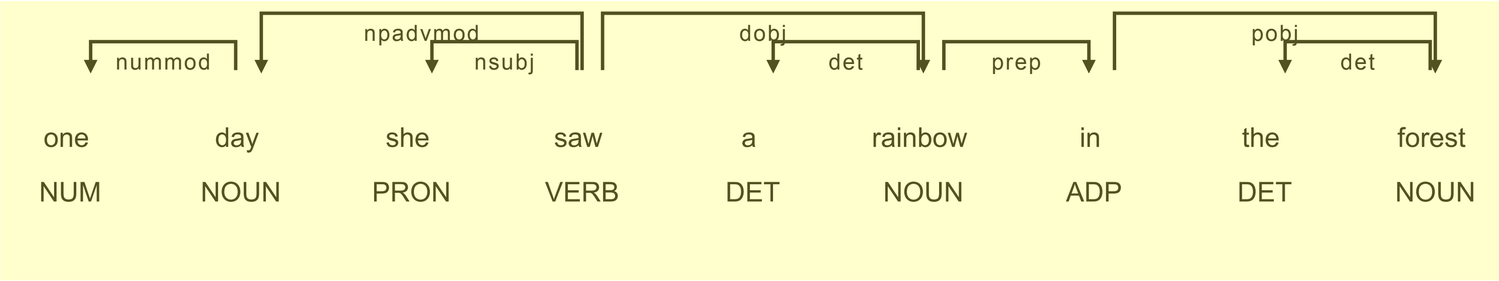


Align words

Transfer tags

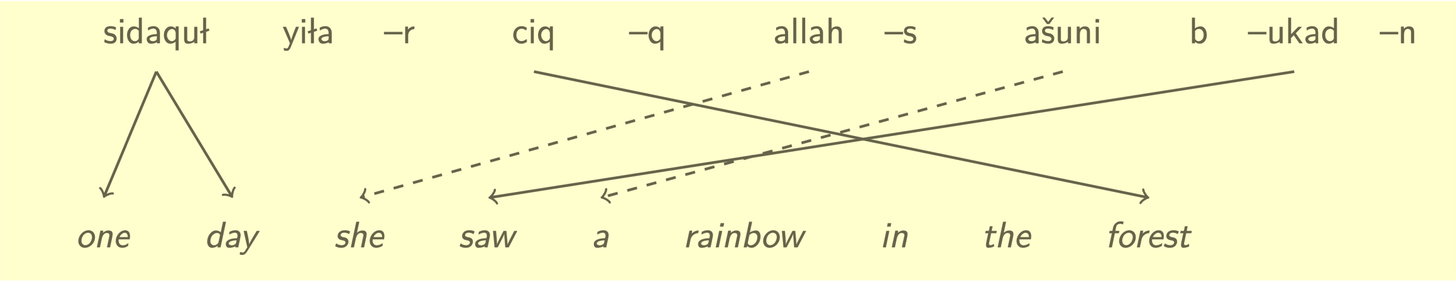


mReps are very useful

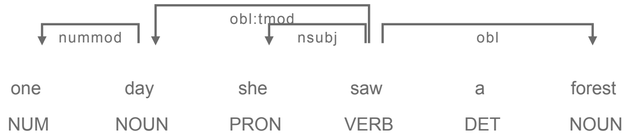


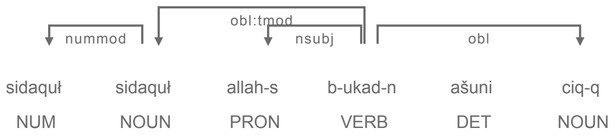
Dependencies for free via alignment links

Parse source

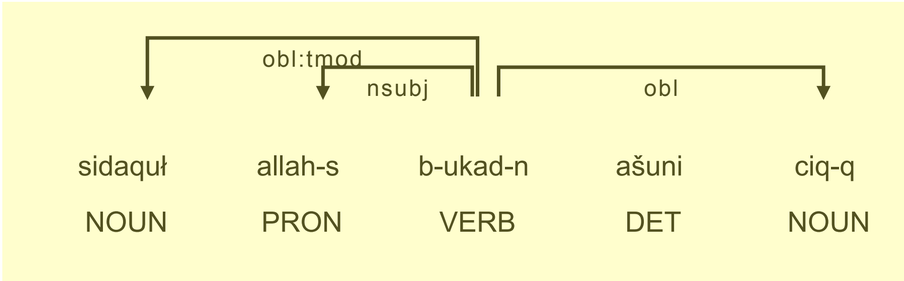


Align words





Transfer links



Multilingual Structural Projection across Interlinear Text (Xia & Lewis, NAACL 2007)

Analyzing mReps

Universal probing [Universal and Independent: Multilingual Probing Framework for Exhaustive Model Interpretation and Evaluation](#) (Serikov et al., BlackboxNLP 2022)

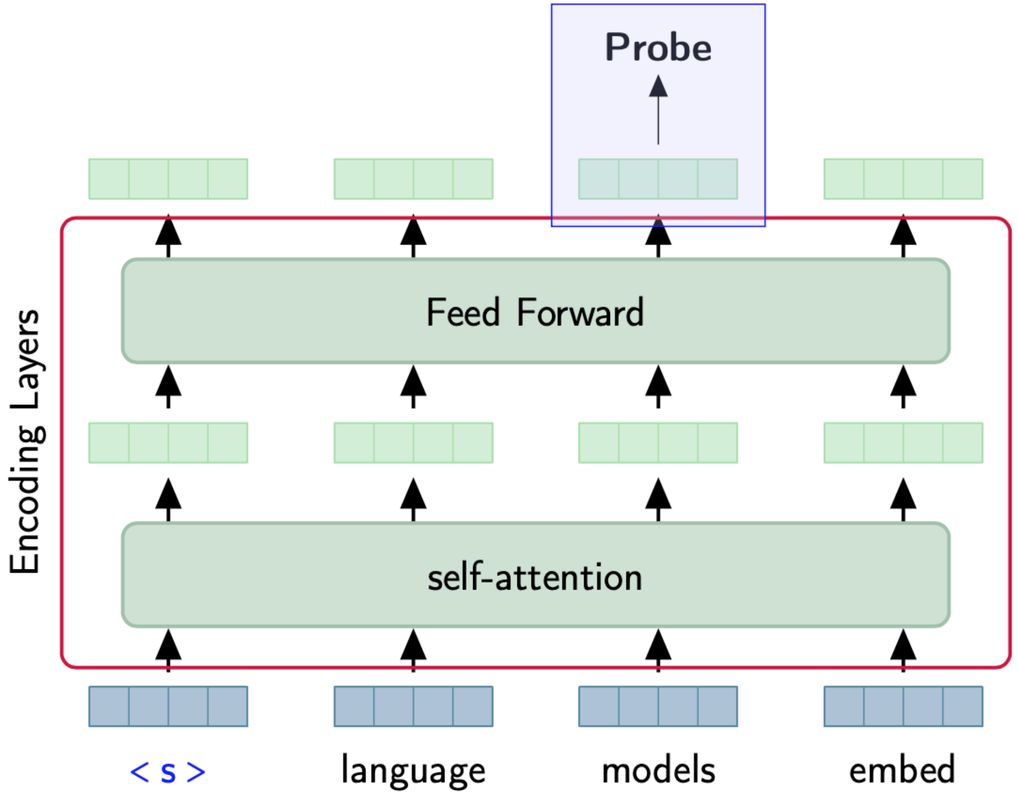
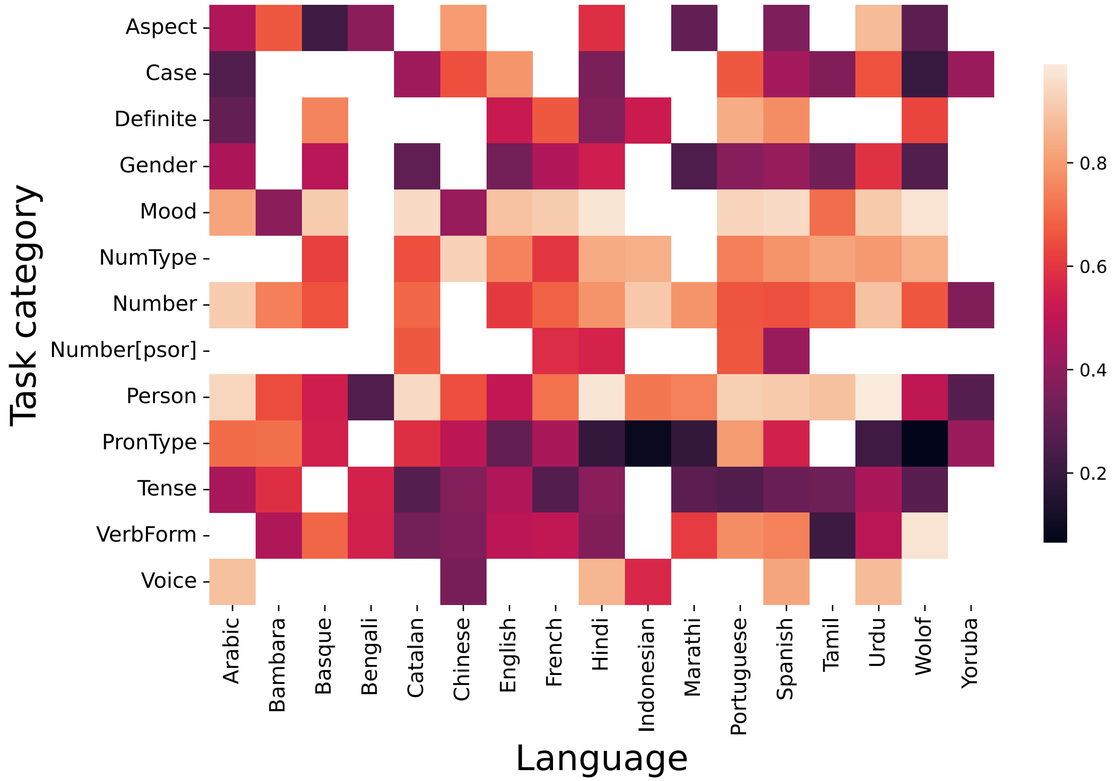


Image (c): G. Wisniewski



[BLOOM: A 176B-Parameter Open-Access Multilingual Language Model](#) (BigScience Workshop, arxiv 2022)

Evaluating mLLMs

mLLMs as monolingual models

mLLMs as representation learners

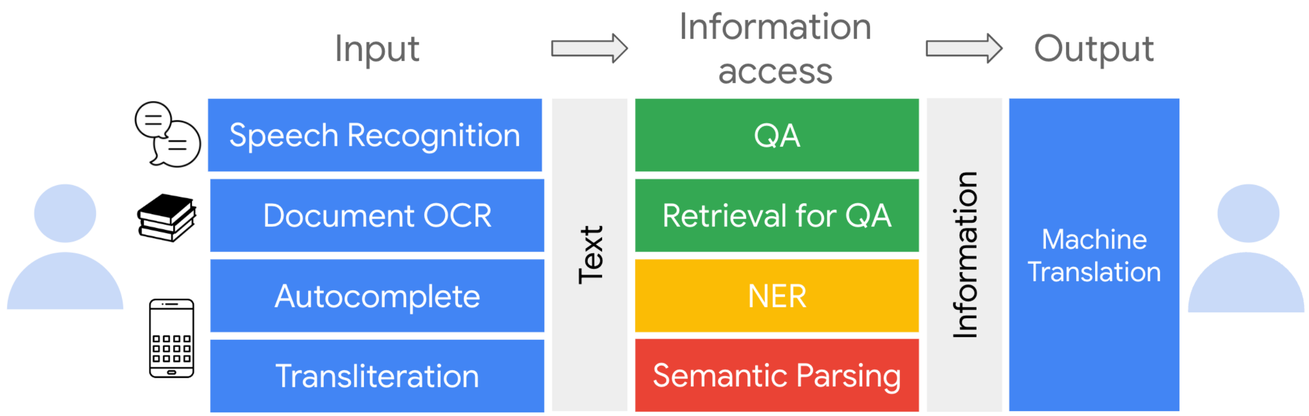
- recovering good bilingual associations
- yielding good (word, sentence) alignments
- encoding linguistic features
- effective X-lingual performance

mLLMs for text generation

- good models of multilingual texts
- good machine translation systems
- discriminating well vs ill-formed sentences
- generating realistic mixed-language ?

A wealth of multilingual benchmarks

- XTreme, X-GLUE, XTreme-R
- Mega, MegaVerse
- BUFFET
- XTreme-Up

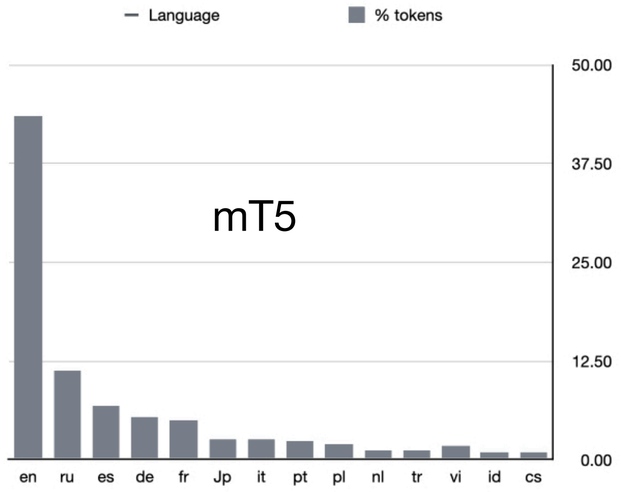


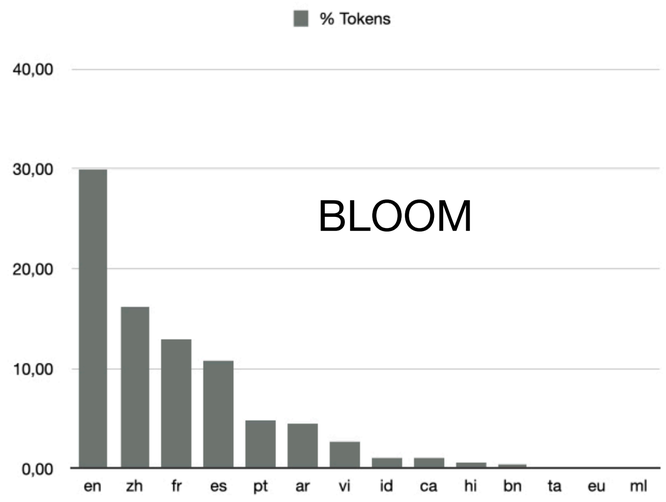
Xtremeup: A User-Centric Scarce-Data Benchmark for Under-Represented Languages (Ruder et al, 2023)

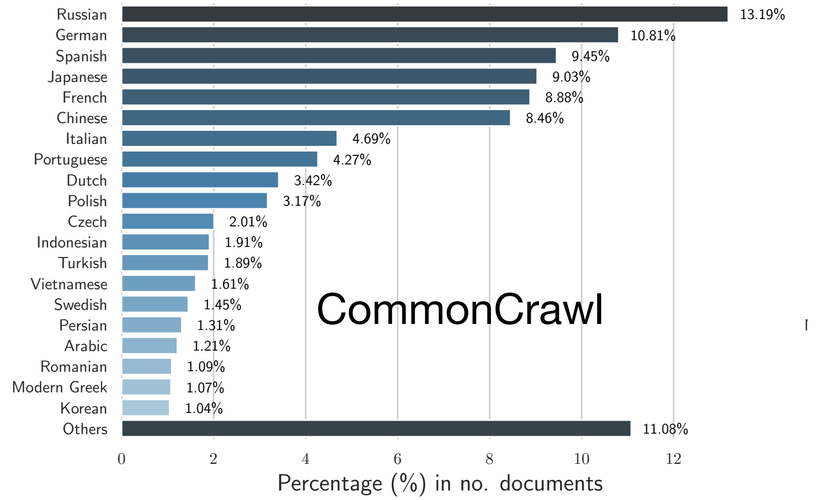
The landscape of mLLMs

2019-2022 - « High-resource / Mid-resource »

- « pure encoder»: mBERT (100 languages), XLM-R (100 languages)
- « pure decoder »: LMs: XGLM (30 languages), Falcon (4+7 languages), BLOOM (46 languages), mGPT (60 languages), PALM (100 languages)
- «enc/dec »: YaLM (Russian-English), GLM (Chinese-English), AlexaTM (12 languages), mBART (50+ languages), mT5 (105 languages),
- multilingual translation models: M2M (100 languages), NLLB (200 languages), etc

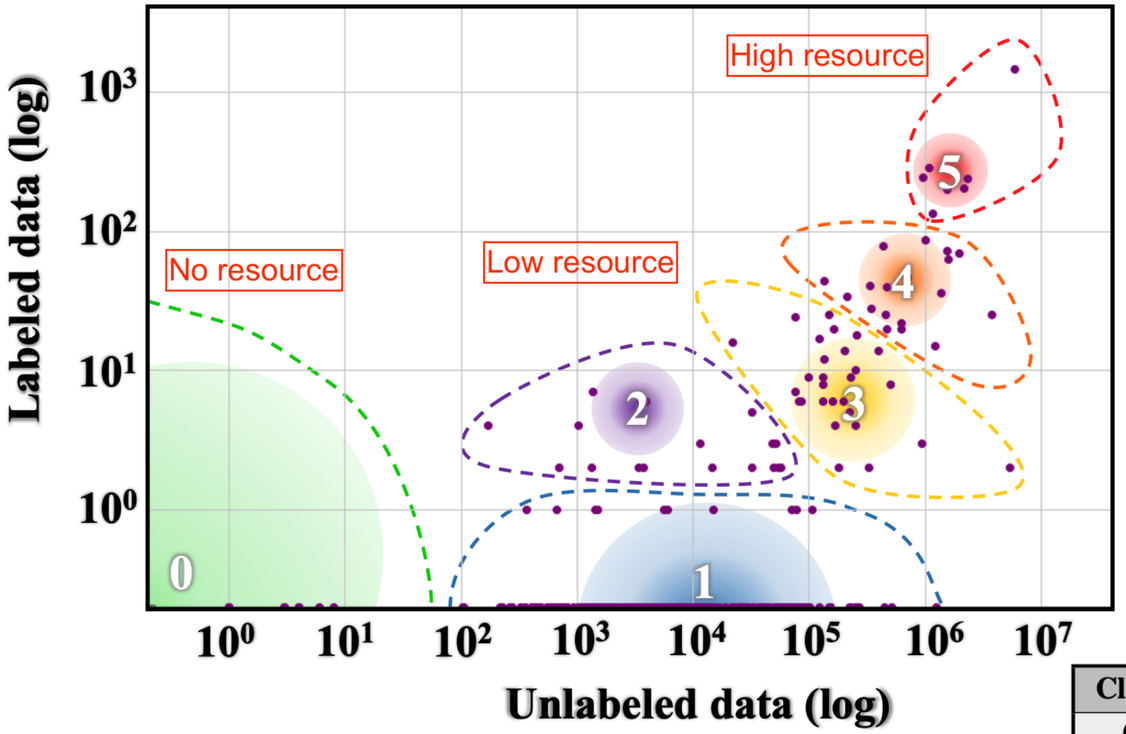






RefinedWeb: The RefinedWeb Dataset for Falcon LLM (Penedo et al, 2023)

Widening the Scope of mLLMs



Main challenge: lack of monolingual data

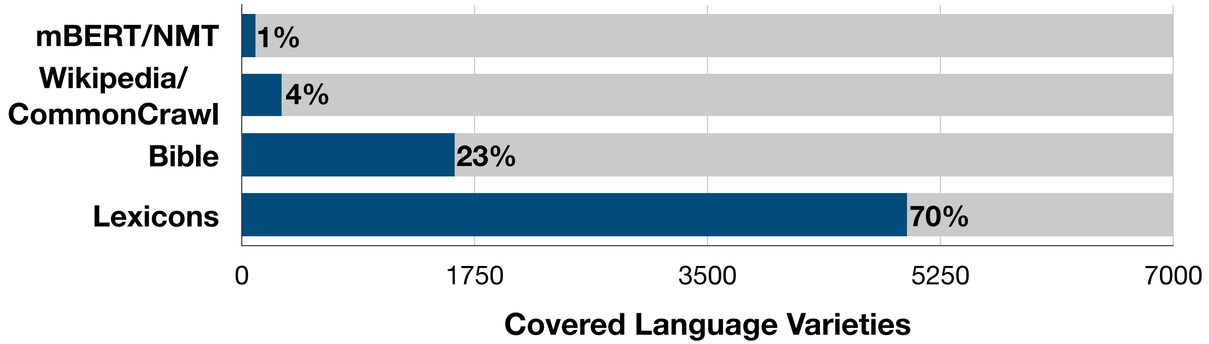
Class	5 Example Languages	#Langs	#Speakers	% of Total Langs
0	Dahalo, Warlpiri, Popoloca, Wallisian, Bora	2191	1.2B	88.38%
1	Cherokee, Fijian, Greenlandic, Bhojpuri, Navajo	222	30M	5.49%
2	Zulu, Konkani, Lao, Maltese, Irish	19	5.7M	0.36%
3	Indonesian, Ukranian, Cebuano, Afrikaans, Hebrew	28	1.8B	4.42%
4	Russian, Hungarian, Vietnamese, Dutch, Korean	18	2.2B	1.07%
5	English, Spanish, German, Japanese, French	7	2.5B	0.28%

The State and Fate of Linguistic Diversity and Inclusion in the NLP World (Joshi et al., ACL 2020)

The landscape of mLLMs

2022-2023 - « widening the scope »

- Artificial data from lexicon [Expanding Pretrained Models to Thousands More Languages via Lexicon-based Adaptation](#) (Wang et al., ACL 2022)



- Search and cure
  - AfriBerta (11 languages), IndicNLPSuite (11 languages), etc
  - Serengeti (517 languages) [SERENGETI: Massively Multilingual Language Models for Africa](#) (Idebara et al, 2023)
  - MadLad-400 (400 languages) [MADLAD-400: A Multilingual And Document-Level Large Audited Dataset](#) (Kudunguta et al, 2023)
  - **Glot500-m** (500 languages) [Glot500: Scaling Multilingual Corpora and Language Models to 500 Languages](#) (ImaniGooghari et al., ACL 2023)

Glot500-c: design choices

1. Select reliable sources

- curated multilingual corpora
- new data crawls
- multiple domains: web, news, science, religion, etc
- excludes toxicity by design

2. Language Identification

- per sentence LID
- joint detection of language + script

3. Chunks and sentence filtering

- high character repetitions
- normal / special char ratios
- insufficient number of words
- wrong language / script
- char LM filtering
- duplicate removal

4. Final selection

- > 30k sentences
- from 2000+ to 511 language / 34 scripts
- 1.5b sentences
- head vs. tail languages
- 1000 dev+test sentences / languages

Glot500-c: the artefact

100 « head »
400 « tail »

Language-Script	|Sent|	Family	Head	Language-Script	|Sent|	Family	Head	Language-Script	|Sent|	Family	Head
hbs_Latn	63411156	indo1319		vec_Latn	514240	indo1319		swh_Latn	95776	atla1278	yes
mal_Mlym	48098273	drav1251	yes	jpn_Jpan	510722	japo1237	yes	alt_Cyrl	95148	turk1311	
aze_Latn	46300705		yes	lus_Latn	509250	sino1245		rmn_Grek	94533	indo1319	
guj_Gujr	45738685	indo1319	yes	crs_Latn	508755	indo1319		miq_Latn	94343	misu1242	
ben_Beng	43514870	indo1319	yes	kqn_Latn	507913	atla1278		kaa_Cyrl	88815	turk1311	
kan_Knda	41836495	drav1251	yes	ndo_Latn	496613	atla1278		kos_Latn	88603	aust1307	
tel_Telu	41580525	drav1251	yes	snd_Arab	488730	indo1319	yes	grn_Latn	87568		
mlt_Latn	40654838	afro1255		yue_Hani	484700	sino1245		lhu_Latn	87255	sino1245	
fra_Latn	39197581	indo1319	yes	tiv_Latn	483064	atla1278		lzh_Hani	86035	sino1245	
spa_Latn	37286756	indo1319	yes	kua_Latn	473535	atla1278		ajp_Arab	83297	afro1255	
eng_Latn	36122761	indo1319	yes	kwy_Latn	473274	atla1278		cmn_Hani	80745	sino1245	yes
fil_Latn	33493255	aust1307	yes	hin_Latn	466175	indo1319		gcf_Latn	80737	indo1319	
nob_Latn	32869205	indo1319		iku_Cans	465011			rmn_Cyrl	79925	indo1319	
rus_Cyrl	31787973	indo1319	yes	kal_Latn	462430	eski1264		kjh_Cyrl	79262	turk1311	
deu_Latn	31015993	indo1319	yes	tdt_Latn	459818	aust1307		rng_Latn	78177	atla1278	
tur_Latn	29184662	turk1311	yes	gsw_Latn	449240	indo1319		mgh_Latn	78117	atla1278	
pan_Guru	29052537	indo1319	yes	mfe_Latn	447435	indo1319		xmv_Latn	77896	aust1307	
mar_Deva	28748897	indo1319	yes	swc_Latn	446378	atla1278		ige_Latn	77114	atla1278	
por_Latn	27824391	indo1319	yes	mon_Latn	437950	mong1349		rmy_Latn	76991	indo1319	
nld_Latn	25061426	indo1319	yes	mos_Latn	437666	atla1278		srm_Latn	76884	indo1319	
ara_Arab	24524122		yes	kik_Latn	437228	atla1278		bak_Latn	76809	turk1311	
zho_Hani	24143786		yes	cnh_Latn	436667	sino1245		gur_Latn	76151	atla1278	
ita_Latn	23539857	indo1319	yes	gil_Latn	434529	aust1307		idu_Latn	75106	atla1278	
ind_Latn	23018106	aust1307	yes	pon_Latn	434522	aust1307		yom_Latn	74818	atla1278	
ell_Grek	22033282	indo1319	yes	umb_Latn	431589	atla1278		tdx_Latn	74430	aust1307	
bul_Cyrl	21823004	indo1319	yes	lvs_Latn	422952	indo1319		mzn_Arab	73719	indo1319	
swe_Latn	20725883	indo1319	yes	sco_Latn	411591	indo1319		cfm_Latn	70227	sino1245	
ces_Latn	20376340	indo1319	yes	ori_Orya	410827		yes	zpa_Latn	69237	otom1299	
isl_Latn	19547941	indo1319	yes	arg_Latn	410683	indo1319		kbd_Cyrl	67914	abkh1242	
pol_Latn	19339945	indo1319	yes	kur_Latn	407169	indo1319	yes	lao_Laoo	66966	taik1256	yes
ron_Latn	19190217	indo1319	yes	dhv_Latn	405711	aust1307		nap_Latn	65826	indo1319	
dan_Latn	19174573	indo1319	yes	luo_Latn	398974	nilo1247		qub_Latn	64973	quec1387	

Usual suspects
Minority script
Minority variety
Minority language

Training Glot500-m

Starting point: XML-R-(B)

- 100 head languages
- « Pure encoder »
- Trained on CommonCrawl with MLM loss
- 250k vocab with *SentencePiece (SP)*
- base (270m) and large (550m) parameters

Unsupervised Cross-lingual Representation Learning at Scale (Conneau et al., ACL 2020)

Training regime

- random language mixtures (temp = 0.3)
- MLM loss
- no change in model size
- two weeks of computations

Extended subword vocabulary

- train SP model on Glot500-m (250k)
- temp = 0.3
- merge « old » and « new » types
- 401k vocabulary

An « academic » configuration

Evaluating Glot500-m

Text classification

- Sentence Classification (90 head, 284 tail) F1 using Taxi1500

Use zero-shot transfer from English
6-way classification for test data (Bible)

Taxi1500: A Multilingual Dataset for Text Classification in 1500 Languages
C Ma et al (2023) arXiv preprint arXiv:2305.08487,

Sentence labeling

- NER (89 head, 75 tail), F1 using wikiAnn
- POS (63 head, 28 tail), F1 using UD

Use zero-shot transfer from English
Requires gold labels

Standard benchmarks with fine grained annotations
Mostly head languages

Evaluating Glot500-m

Sentence Retrieval

- Tatoeba (68 head, 28 tail), acc@10
- Bible (94 head, 276 tail), acc@10

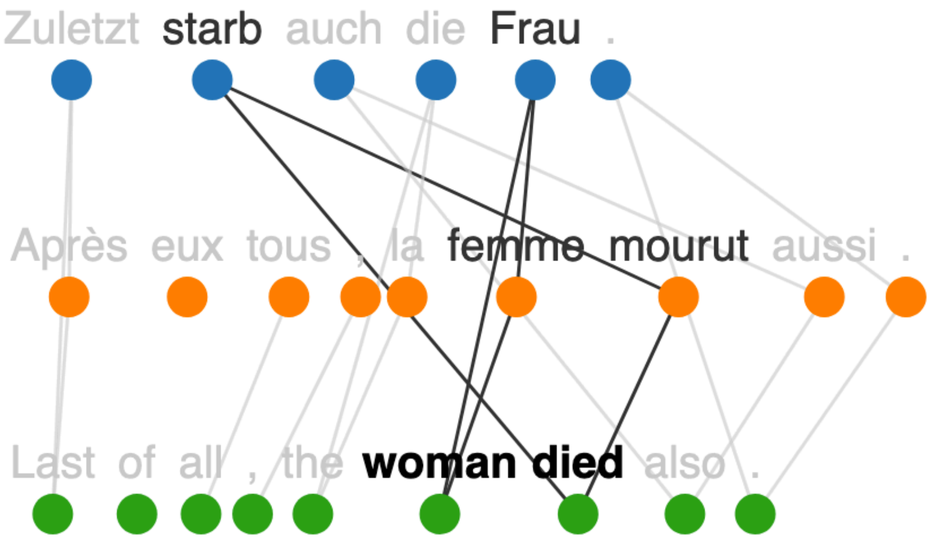
Find nearest foreign neighbor in multilingual space for 1000 (resp. 500) English sentences
Requires parallel data

Unsupervised multilingual evaluation

Round trip Alignment

- Bible (95 head, 288 tail), acc@10

Word *SimAlign* $L_1 \rightarrow L_2 \rightarrow L_3 \rightarrow L_4 \rightarrow L_1$
Report exact matches
Requires parallel data



Improving tail languages

	all			head			tail		
	XLM-R-B	XLM-R-L	Glot500-m	XLM-R-B	XLM-R-L	Glot500-m	XLM-R-B	XLM-R-L	Glot500-m
TextClass	23.3	25.8	**48.7**	51.3	**60.5**	54.7	13.7	13.9	**46.6**
NER	55.3	59.5	**62.4**	61.8	**66.0**	63.9	47.5	51.8	**60.7**
POS	65.8	67.7	**71.8**	76.4	**78.4**	76.0	41.7	43.5	**62.3**
SR [Tatoeba]	56.6	60.4	**70.7**	66.2	71.1	**75.0**	32.6	33.6	**59.8**
SR [Bible]	19.3	20.1	**47.3**	54.2	58.3	**59.0**	7.4	7.1	**43.2**
RTA	2.8	3.3	**4.7**	3.4	4.1	**5.5**	2.6	3.1	**4.5**

Glot500-m vs XML

- outperforms all models on average
- better than XML-R-B for head languages
- much better than XML-R-* for tail languages

Also

- are averages that informative ?
- tail language scores remain poor

Check paper for complete results / language

Where Glot500-m really helps

		language-script	XLMR	Glot500	gain		language-script	XLMR	Glot500	gain
high end	SentRetr Tatoeba	tat C Tatar	10.3	70.3	60.0	SentRetr Bible	uzn C Northern Uzbek	5.4	87.0	81.6
		nds L Low German	28.8	77.1	48.3		crs L Seselwa Creole	7.4	80.6	73.2
		tuk L Turkmen	16.3	63.5	47.3		srn L Sranan Tongo	6.8	79.8	73.0
		ile L Interlingue	34.6	75.6	41.0		uzb C Uzbek	6.2	78.8	72.6
		uzb C Uzbek	25.2	64.5	39.3		bcl L Central Bikol	10.2	79.8	69.6
low end		dtp L Kadazan Dusun	5.6	21.1	15.5		xav L Xavánte	2.2	5.0	2.8
		kab L Kabyle	3.7	16.4	12.7		mau L Huautla Mazatec	2.4	3.6	1.2
		pam L Pampanga	4.8	11.0	6.2		ahk L Akha	3.0	3.2	0.2
		lvs L Standard Latvian	73.4	76.9	3.5		aln L Gheg Albanian	67.8	67.6	-0.2
		nob L Bokmål	93.5	95.7	2.2		nob L Bokmål	82.8	79.2	-3.6
high end	NER	div T Dhivehi	0.0	50.9	50.9	POS	mlt L Maltese	21.3	80.3	59.0
		che C Chechen	15.3	61.2	45.9		sah C Yakut	21.9	76.9	55.0
		mri L Maori	16.0	58.9	42.9		sme L Northern Sami	29.6	73.6	44.1
		nan L Min Nan	42.3	84.9	42.6		yor L Yoruba	22.8	64.2	41.4
		tgk C Tajik	26.3	66.4	40.0		quc L K'iche'	28.5	64.1	35.6
low end		zea L Zeeuws	68.1	67.3	-0.8		lzh H Literary Chinese	11.7	18.4	6.7
		vol L Volapük	60.0	59.0	-1.0		nap L Neapolitan	47.1	50.0	2.9
		min L Minangkabau	42.3	40.4	-1.8		hyw A Western Armenian	79.1	81.1	2.0
		wuu H Wu Chinese	28.9	23.9	-5.0		kmr L Northern Kurdish	73.5	75.2	1.7
		lzh H Literary Chinese	15.7	10.3	-5.4		aln L Gheg Albanian	54.7	51.2	-3.5

- Scripts not in XMLR
- Large training data
- Cluster effects

- Neighbours in XMLR

Also: the « curse of multilinguality » - myth or reality ?

Take away

A useful artefact

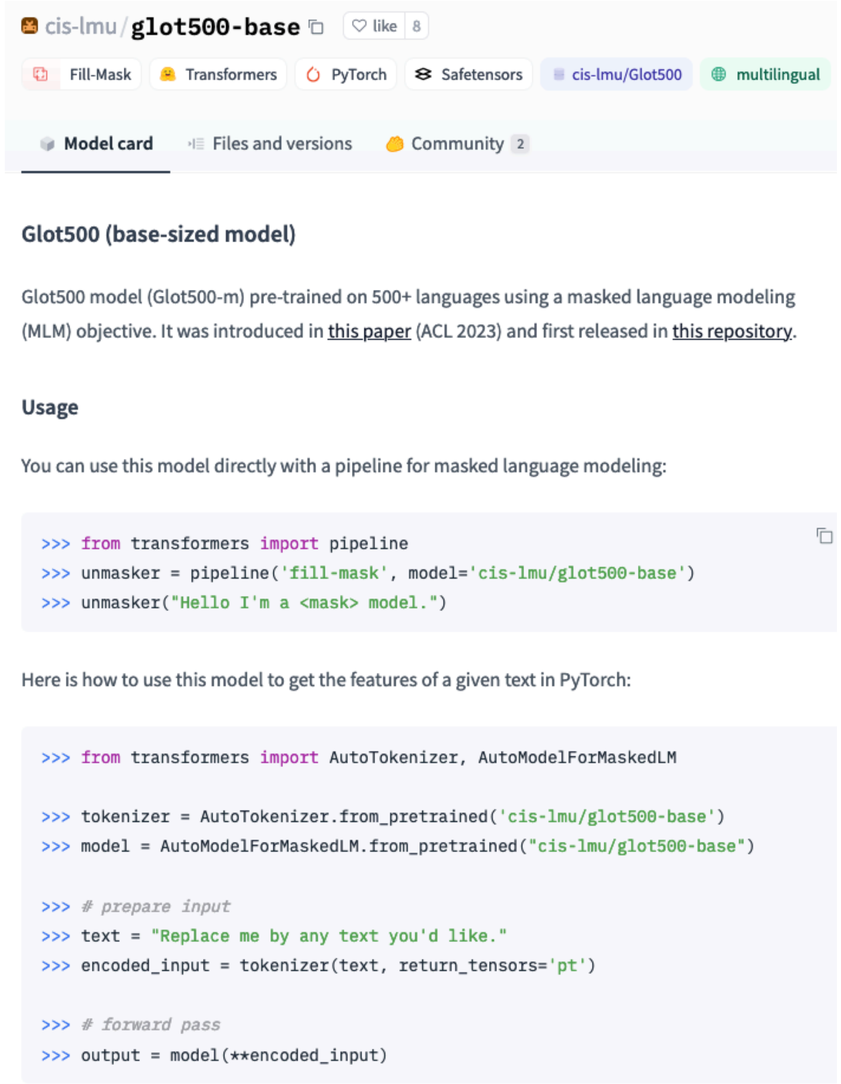
- extends the scope of « modern » NLP for tail languages
- open-source, documented corpus collection

To be continued

- deeper analysis of results
- analysis by langage families
- linguistic analysis of representations
- evaluation of text generation

Limitations

- corpus selection and curation
- subword vocab size and training
- language choices
- filtering strategies



https://huggingface.co/cis-lmu/glot500-base

Towards 1000 languages

2. Language Identification

- per sentence LID
- joint detection of language + script

Is this reliable ? Will it scale ?

Language Identification

Many tools

- FastText LID
- Compact Language Detector (CLD2, CLD3)
- and more (whatlang, franc, idNet, openLID…)

Mostly using ML techniques

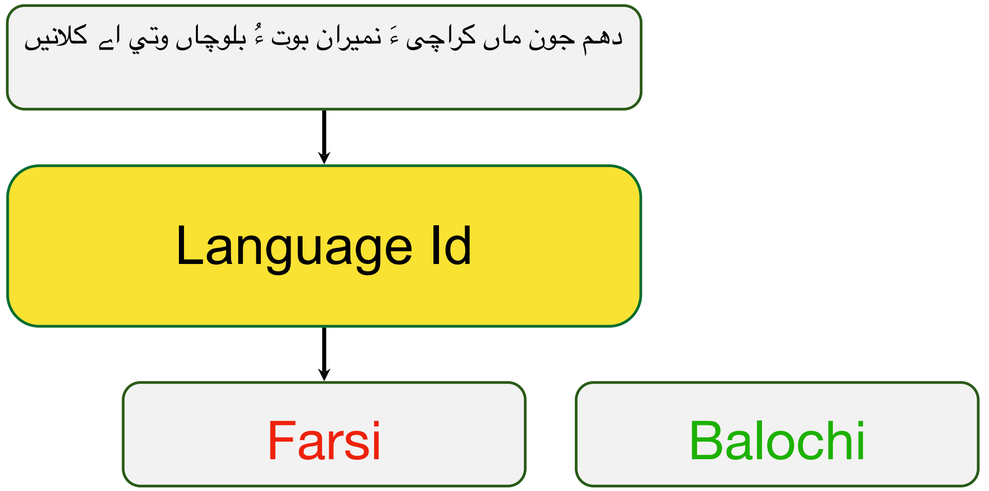
NB, SVM, LogReg, NNets

A solved problem ?

Language Identification at scale

Existing tools are limited

Fastext LID	176
Compact Langage Detector	107
whatlang	69
OpenLID	200
franc-s	82
franc-m	187
franc-l	417



« out-of-model cousin error »

Other issues

- speed
- implementation / deployment
- lack of openness
- lack of documentation
- lack confidence estimation
- lack of rejection model

Designing your own

- selecting / naming languages
- curating corpora
- the language mix, again

Language Identification

ISO 639-3 (3 letters) also ISO 639-1 (2 letters) 7000+

gsw Alsatian	cat Catalan	fsl French Sign Language	oci Occitan
frp Arpitan	cos Corsican	lij Ligurian	pcd Picard
eus Basque	emx Erromintxela	pfl Lorraine Franconian	sdt Shuadit
bre Breton	fra French	ltz Luxembourgish	vls West Flemish

Ethnologue:https://www.ethnologue.com/country/FR/

BCP 47 language-extlang-script-region-variant-extension-privateuse

Code	Language	Subtags
en	English	language
mas	Maasai	language
fr-CA	French as used in Canada	language+region
es-419	Spanish as used in Latin America	language+region
zh-Hans	Chinese written with Simplified script	language+script

Training GlotLID

Train corpus

- >1800+ languages scripts
- multiple reliable sources
- mixture of domains:
  - Wikipedia
  - religious texts
  - collaborative translations
  - academia
  - storybooks
  - news sites

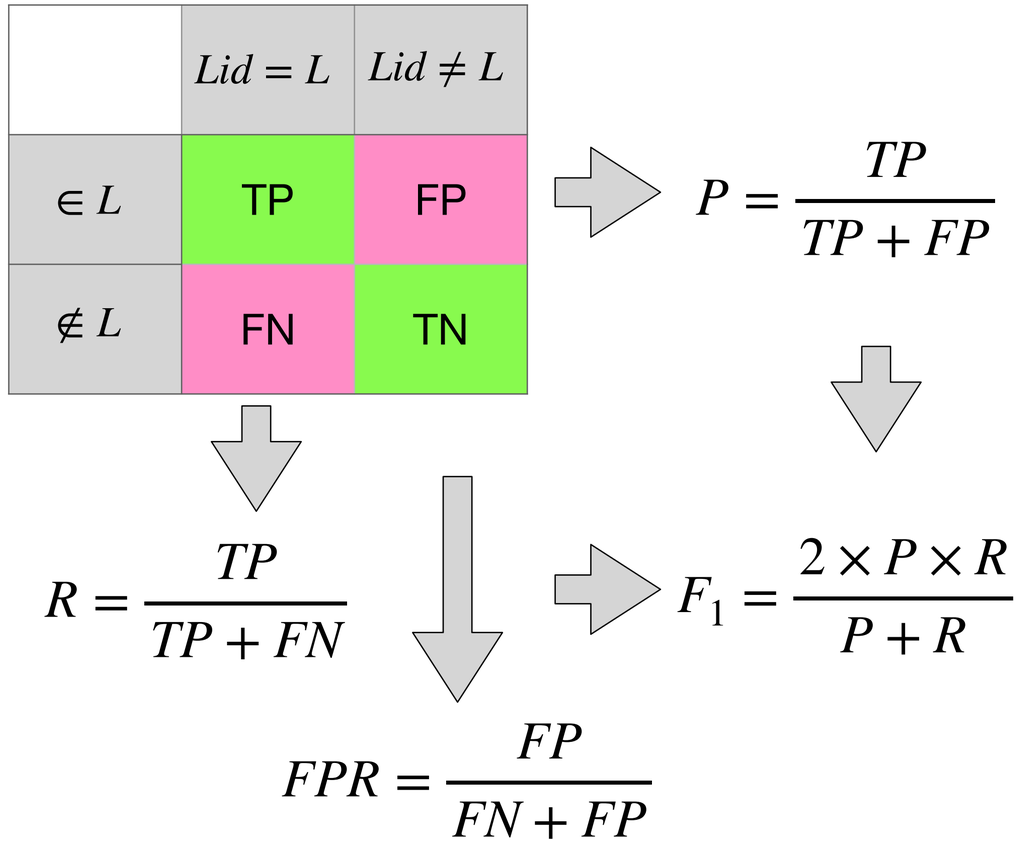
Implementation: FastText

- linear classifier
- char & word n-gram features
- highly optimized for speed (training and inference)

Bag of Tricks for Efficient Text Classification (Joulin et al., EACL 2017)

Evaluating LID

- GlotLID-C (testset)
  - 1832 language-scripts
- FLORES-200
  - 204 distinct language-scripts
  - human verified translations
- Universal Declaration of Human Rights
  - exist in 500+ languages
  - partial overlap with Wikipedia

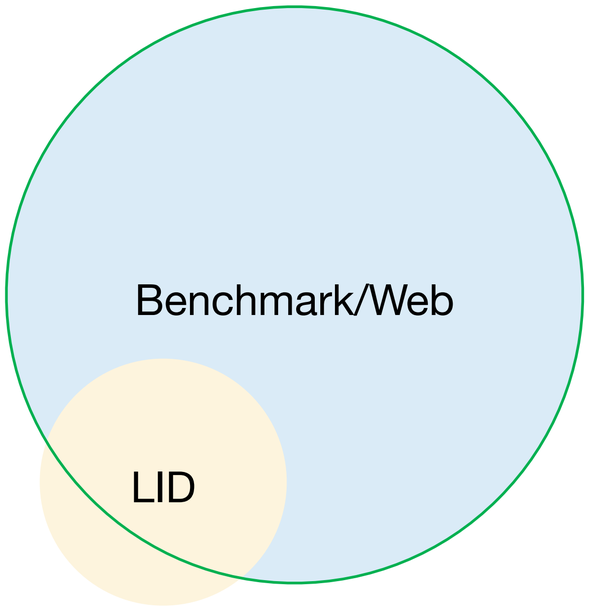


Evaluating GlotLID

- rejection

+ rejection

Benchmark		|L|	GlotLID-M, θ=.0 F1↑	GlotLID-M, θ=.0 FPR↓	GlotLID-M, θ=.5 F1↑	GlotLID-M, θ=.5 FPR↓
GlotLID-C	all	1832	.940	.0005	.938	.0003
GlotLID-C	subset	1665	.977	.0003	.973	.0002
UDHR	all	374	.750	.0015	.734	.0007
UDHR	subset	342	.784	.0014	.770	.0006
FLORES-200	all	196	.917	.0042	.887	.0013
FLORES-200	subset	177	.957	.0029	.924	.0010



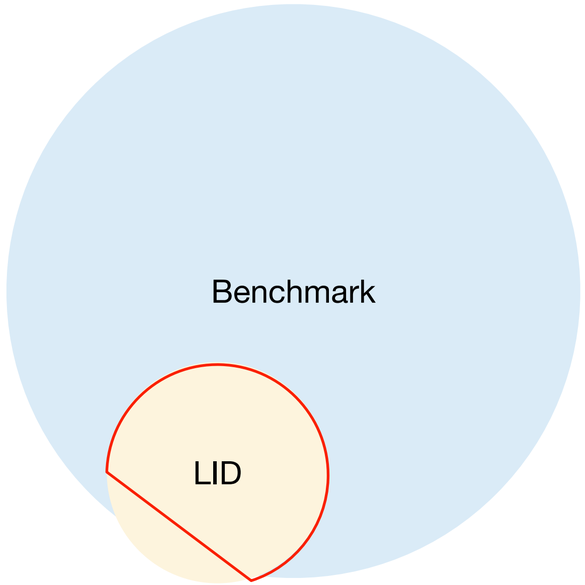
« Realistic setting »

- set of possible languages is unknown
- rejection matters

Language verification

Closed language assumption

- set of possible languages known (differs for each baseline)



		FLORES-200								UDHR							
		CLD3		FT176		OpenLID		NLLB		CLD3		FT176		OpenLID		NLLB	
		$\|L\| = 96$		$\|L\| = 108$		$\|L\| = 195$		$\|L\| = 188$		$\|L\| = 100$		$\|L\| = 124$		$\|L\| = 159$		$\|L\| = 172$	
LID Model	θ	**F1↑**	**FPR↓**	**F1↑**	**FPR↓**	**F1↑**	**FPR↓**	**F1↑**	**FPR↓**	**F1↑**	**FPR↓**	**F1↑**	**FPR↓**	**F1↑**	**FPR↓**	**F1↑**	**FPR↓**
baselines	.0	.952	**.0104**	.881	**.0093**	**.923**	**.0051**	.950	**.0053**	.922	.0101	.739	**.0081**	.881	**.0063**	.854	**.0058**
GlotLID-M	.0	**.983**	**.0104**	**.991**	**.0093**	.922	**.0051**	**.954**	**.0053**	**.952**	**.0100**	**.927**	**.0081**	**.926**	.0064	**.925**	.0060

Corpus building for low-resource languages

- Negatives >> Positives
- FPR matters
- F1 not important for HR

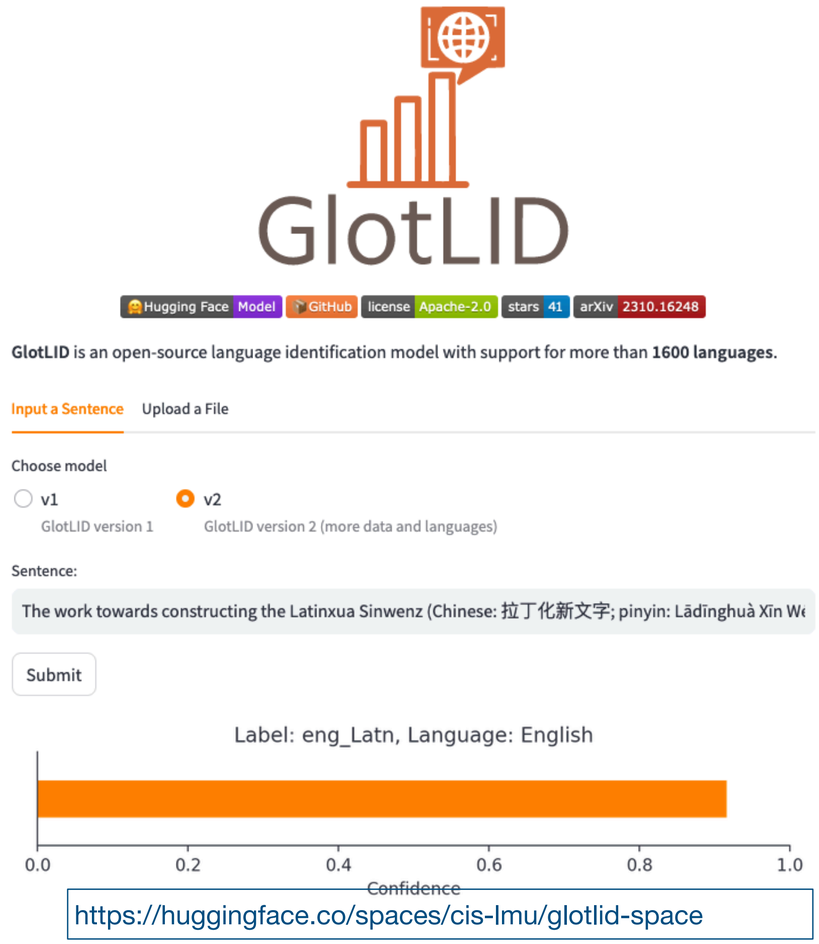
Language Identification unsolved

	FLORES-200					UDHR				
	language	FP	cl	top FP source	#FP	%	language	FP	cl	top FP source	#FP	%
most errors	arb:St Arabic	3787	.18	ars:Najdi Arabi	829	.22	cmn:Mandarin Ch	596	.38	chr:Cherokee	81	.14
	arz:Egyptian Ar	1726	.32	apc:Levantine A	440	.25	qub:Huallaga Hu	247	.00	qvh:Huamalíes-D	55	.22
	pes:Ir. Persian	1495	.40	prs:Dari	905	.61	fin:Finnish	224	.22	krl:Karelian	138	.62
	cmn:Mandarin Ch	1008	.00	yue:Yue Chinese	1008	.99	wuu:Wu Chinese	172	.24	hak:Hakka Chine	44	.26
	hin:Hindi	977	.51	awa:Awadhi	693	.71	rus:Russian	157	.28	niv:Gilyak	44	.28
most noisy	arb:St Arabic	3787	.18	ars:Najdi Arabi	829	.22	evn:Evenki	36	.23	oaa:Orok	19	.53
	arz:Egyptian Ar	1726	.32	apc:Levantine A	440	.25	quz:Cusco Quech	82	.40	qxu:Arequipa-La	61	.74
	prs:Dari	338	.24	pbt:S Pashto	310	.92	hrv:Croatian	84	.42	bos:Bosnian	39	.46
	dyu:Dyula	255	.25	bam:Bambara	255	.99	tzm:C Atlas Tam	52	.02	zgh:St Moroccan	52	.99
	apc:Levantine A	161	.42	ajp:S Levantine	70	.43	uzn:N Uzbek	72	.46	cbu:Candoshi-Sh	16	.22

Kargaran et al 2023: https://arxiv.org/abs/2310.16248

New issues

- more realistic data
- code-mixed inputs
- improved calibration



https://huggingface.co/spaces/cis-lmu/glotlid-space

Script Identification is easy

Scripts names

- normalized in ISO-15924
- deterministically matched in Unicode 15.0 tables
- 161 scripts documented

Script - language associations

Normalize and aggregate:

- VanEsch et al 2023
- Wikipedia
- ScriptSource (SIL)
- LangTag
- Misc. sources

Alphabets occidentaux modernes [modifier | modifier le code]

	+00	+10	+20	+30	+40	+50	+60	+70	+80	+90	+A0	+B0	+C0	+D0	+E0	+F0
U+0000	commandes C0				latin de base				commandes C1		latin – 1					
U+0100	latin étendu – A								latin étendu – B							
U+0200	latin étendu – B						alphabet phonétique international					lettres modificatives avec chasse				
U+0300	diacritiques							grec et copte								
U+0400	cyrillique															
U+0500	cyrillique – supplément				arménien											

Abjads afroasiatiques modernes (écrits de droite à gauche) [modifier | modifier le code]

	+00	+10	+20	+30	+40	+50	+60	+70	+80	+90	+A0	+B0	+C0	+D0	+E0	+F0
U+0500										hébreu						
U+0600	arabe															
U+0700	syriaque					arabe – supplément			thâna				n'ko			
U+0800	samaritain				mandéen	syriaque – supplément		arabe étendu – B			arabe étendu – A					

Abugidas sud-asiatiques modernes [modifier | modifier le code]

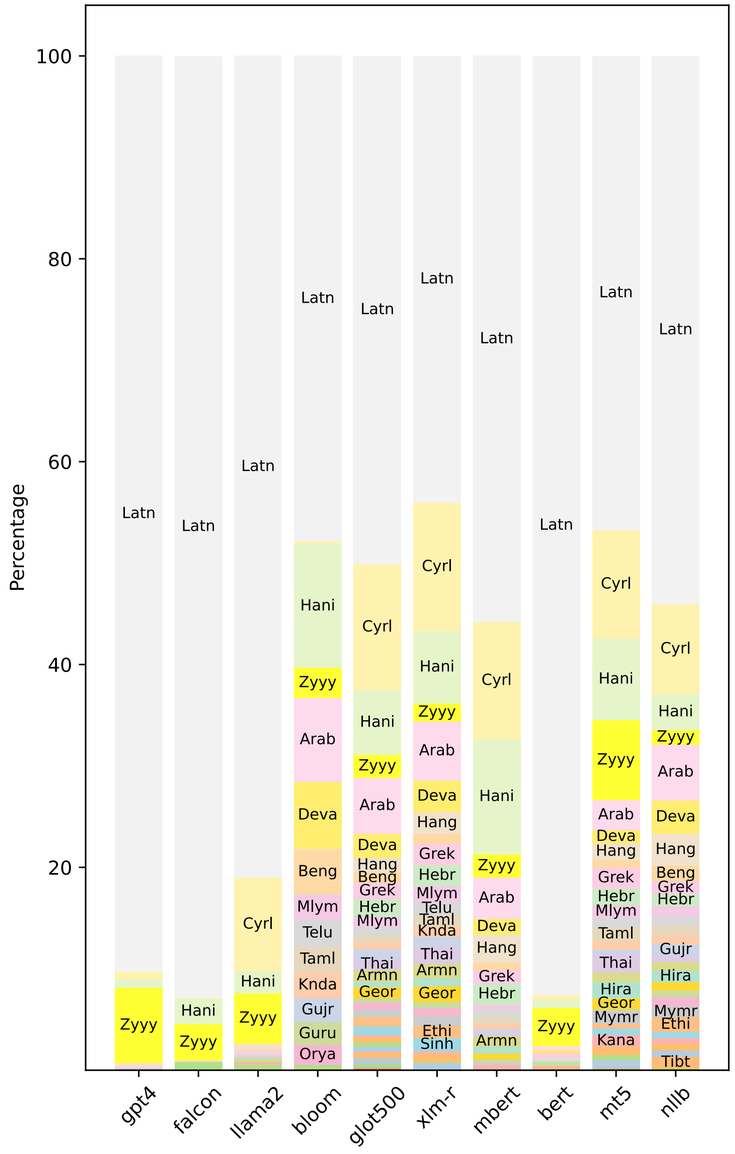
	+00	+10	+20	+30	+40	+50	+60	+70	+80	+90	+A0	+B0	+C0	+D0	+E0	+F0
U+0900	devanâgarî								bengalî							
U+0A00	gourmoukhî (ou gurmukhi)								goudjarati (ou gujarâtî)							
U+0B00	oriyâ (ou odia)								tamoul							
U+0C00	télougou								kannara (ou kannada)							
U+0D00	malayalam								singhalais (ou cingalais)							
U+0E00	thaï								lao (ou laotien)							
U+0F00	tibétain															
U+1000	birman															

Auditing data with GlotScript

		Corpus Code: ISO 639-3	Scripts	ACC↑	ACC70↑	ACC50↑
Highest ACC	mC4	st:sot (S Sotho)	Latn:1000	**1.000**	**1.000**	**1.000**
		fil:fil (Filipino)	Latn:998, Cyrl:1, Hani:1	0.998	0.999	**1.000**
		ro:ron (Romanian)	Latn:996, Zyyy:4, Cyrl:1	0.995	0.997	**1.000**
		id:ind (Indonesian)	Latn:995, Zyyy:3, Hani:1, Hebr:1	0.995	**1.000**	**1.000**
		sw:swa (Swahili)	Latn:995, Zyyy:5	0.995	**1.000**	**1.000**
Lowest ACC		ne:nep (Nepali)	Deva:609, Hani:219, Latn:88, Hang:44, Thai:12, Laoo:8, Zyyy:8, Orya:7, Other:5	0.609	0.730	**0.797**
		mn:mon (Mongolian)	Cyrl:502, Hebr:348, Latn:135, Zyyy:14, Hani:1	0.502	0.557	**0.570**
		cy:cym (Welsh)	Grek:603, Latn:367, Zyyy:11, Hebr:9, Cyrl:5, Zzzz:4, Arab:1	**0.367**	0.338	0.295
		sd:snd (Sindhi)	Latn:654, Arab:329, Zyyy:12, Zzzz:2, Cyrl:1, Hang:1, Telu:1	**0.329**	0.271	0.222
		mr:mar (Marathi)	Hani:454, Thai:252, Latn:119, Deva:116, Zyyy:34, Guru:10, Beng:4, Khmr:3, Other: 8	0.116	0.136	**0.141**
Highest ACC	OSCAR	id:ind (Indonesian)	Latn:998, Zyyy:2	0.998	**1.000**	**1.000**
		war:war (Waray)	Latn:997, Zyyy:3	**0.997**	**0.997**	0.996
		als:gsw (Swiss G)	Latn:996, Zyyy:3, Cyrl:1	0.996	0.996	**1.000**
		vo:vol (Volapük)	Latn:994, Arab:4, Cyrl:1	0.994	**1.000**	**1.000**
		nds:nds (Low G)	Latn:994, Zyyy:2, Cyrl:2, Hang:1, Thaa:1	0.994	**1.000**	**1.000**
Lowest ACC		am:amh (Amharic)	Ethi:822, Latn:164, Zyyy:12, Hani:1, Arab:1	0.822	0.883	**0.940**
		gu:guj (Gujarati)	Gujr:802, Latn:180, Zyyy:12, Deva:6	0.802	0.863	**0.883**
		si:sin (Sinhala)	Sinh:801, Latn:188, Zyyy:11	0.801	0.905	**0.948**
		th:tha (Thai)	Thai:800, Latn:181, Zyyy:18, Hani:1	0.800	0.883	**0.917**
		te:tel (Telugu)	Telu:799, Latn:188, Zyyy:9, Deva:3, Cyrl:1	0.799	0.880	**0.908**

Script-languages mismatches detection finds many errors

Auditing tokenizers



Script Identification Can Help

Kargaran et al, 2023: https://arxiv.org/abs/2309.13320

```
from GlotScript import get_script_predictor
sp = get_script_predictor()
```

```
sp('これは日本人です')
>> ('Hira', 0.625, {'details': {'Hira': 0.625, 'Hani': 0.375}, 'tie': False, 'interval': 0.25})
```

```
sp('This is Latin')[:1]
>> ('Latn', 1.0)
```

```
sp('මේක සිංහල')[0]
>> 'Sinh'
```

```
sp('𝄞𝄫  𒁁𒀸')
>> ('Xsux', 0.5, {'details': {'Xsux': 0.5, 'Zyyy': 0.5}, 'tie': True, 'interval': 0.0})
```

https://github.com/cisnlp/GlotScript

Script identification can help

Kargaran et al, 2023: https://arxiv.org/abs/2309.13320

```
from GlotScript import get_script_predictor
sp = get_script_predictor()
```

```
sp('これは日本人です')
>> ('Hira', 0.625, {'details': {'Hira': 0.625, 'Hani': 0.375}, 'tie': False, 'interval': 0.25})
```

```
sp('This is Latin')[:1]
>> ('Latn', 1.0)
```

```
sp('මේක සිංහල')[0]
>> 'Sinh'
```

```
sp('𝄞𝄫 [illegible]𒀸')
>> ('Xsux', 0.5, {'details': {'Xsux': 0.5, 'Zyyy': 0.5}, 'tie': True, 'interval': 0.0})
```

https://github.com/cisnlp/GlotScript

mLLMs are still a wonder

Open question and issues

mLLMs: more than a collection of monolingual models ?

- which linguistic properties help / break transfer?
- how to measure positive / negative interference?
- how to measure language coverage?

How about the « curse of multilinguality »?

- Impact of language distributions ?
- Impact of model size?
- Impact of vocabulary size?
- How to achieve fairness in mLLMs design?

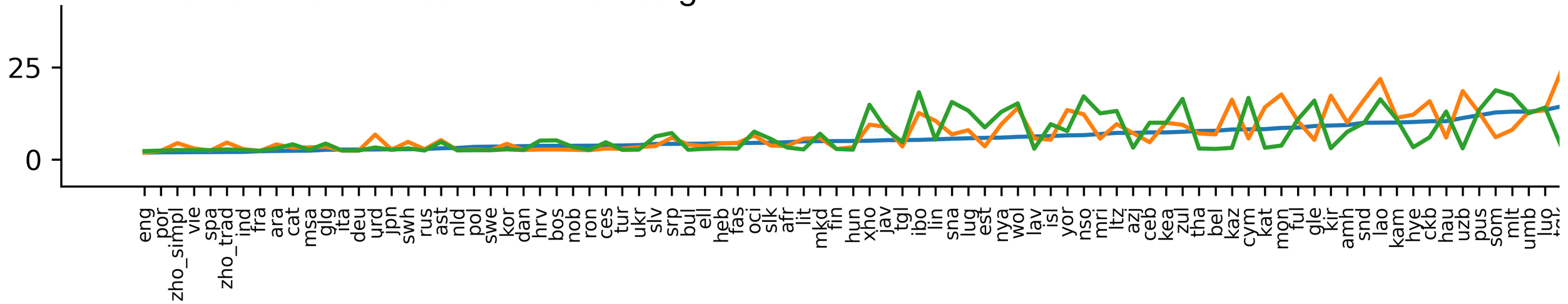
More questions

How about actual multilingual tasks?

- machine translation
- generating code-switched language
- summarization from multilingual texts

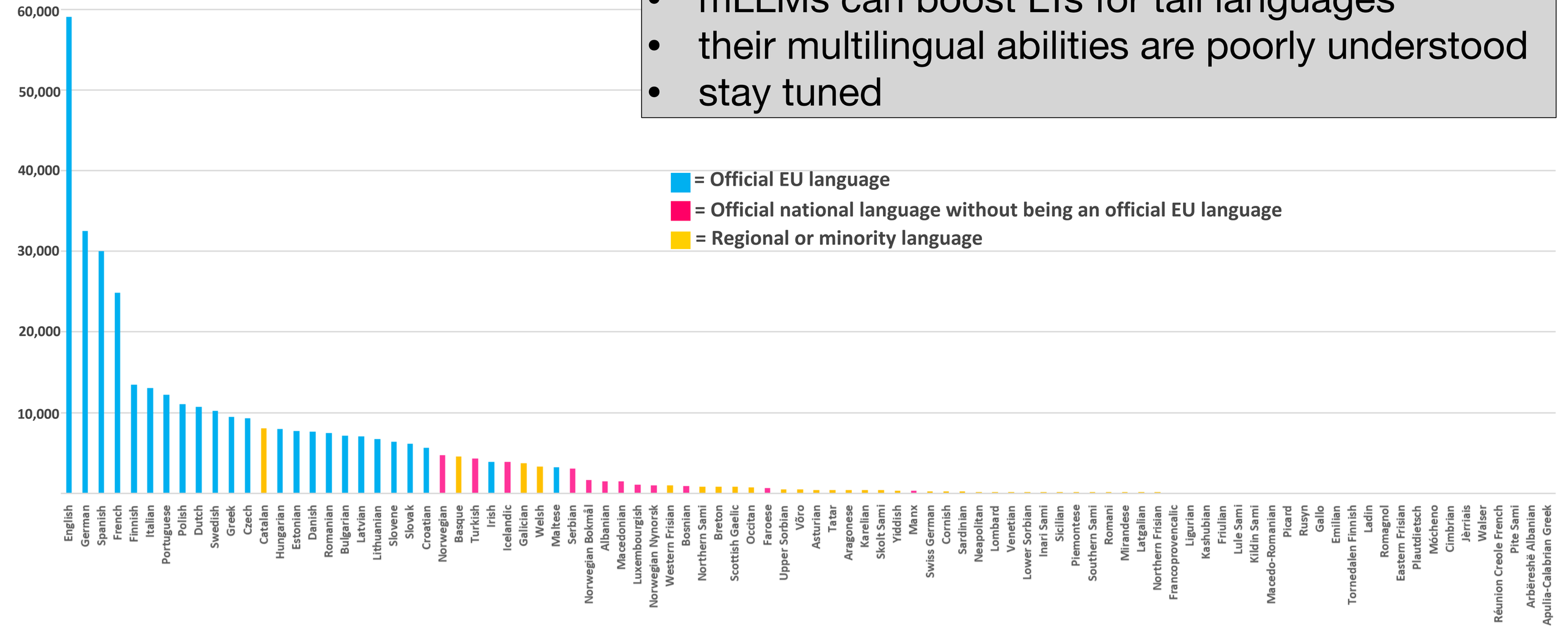
What we need to do

- better models of heterogeneous data
- more diverse samples of language use
- rethink evaluation and scores



Conclusions & Take aways

- mLLMs can boost LTs for tail languages
- their multilingual abilities are poorly understood
- stay tuned



Special thanks

Ayyoob ImaniGooghari
Masoud Jalili Sabet
Amir Hossein Kargaran
Nora Kassner
Lütfi Kerem Şenel
Peiqin Lin

Chunlan Ma
André Martins
Silvia Severini
Helmut Schmid
Hinrich Schütze