GlotCC: An Open Broad-Coverage CommonCrawl Corpus and Pipeline for Minority Languages

Amir Kargaran, François Yvon, Hinrich Schuetze

We present GlotCC, a **clean**, **document-level**, 2TB general domain corpus derived from **CommonCrawl** , covering more than **1000 languages**. We make GlotCC and the system used to generate it— including the pipeline, language identification model, and filters— available to the research community.

Background

The need for large text corpora has **increased** with the advent of pretrained language models, and particularly the discovery of scaling laws for these models.

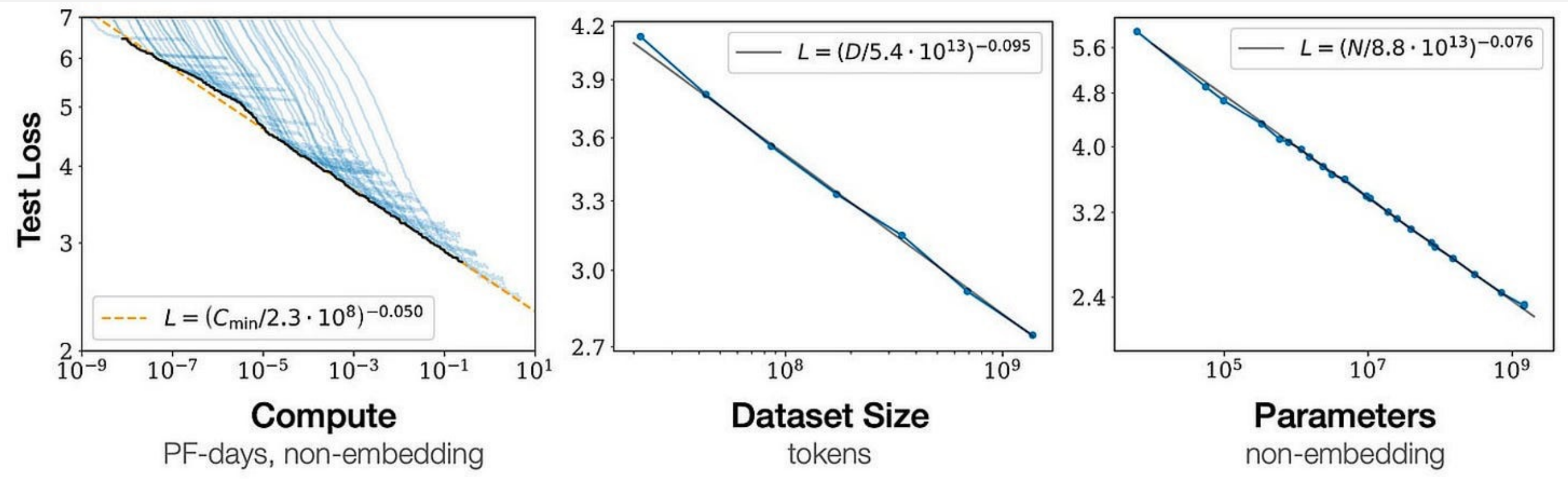


Figure 1 Language modeling performance improves smoothly as we increase the model size, datasetset size, and amount of compute[2] used for training. For optimal performance all three factors must be scaled up in tandem. Empirical performance has a power-law relationship with each individual factor when not bottlenecked by the other two.

Source: Scaling Laws for Neural Language Models, 2020

Most available corpora have sufficient data only for languages with **large dominant communities**.

Source: The State and Fate of Linguistic Diversity and Inclusion in the NLP World, 2020

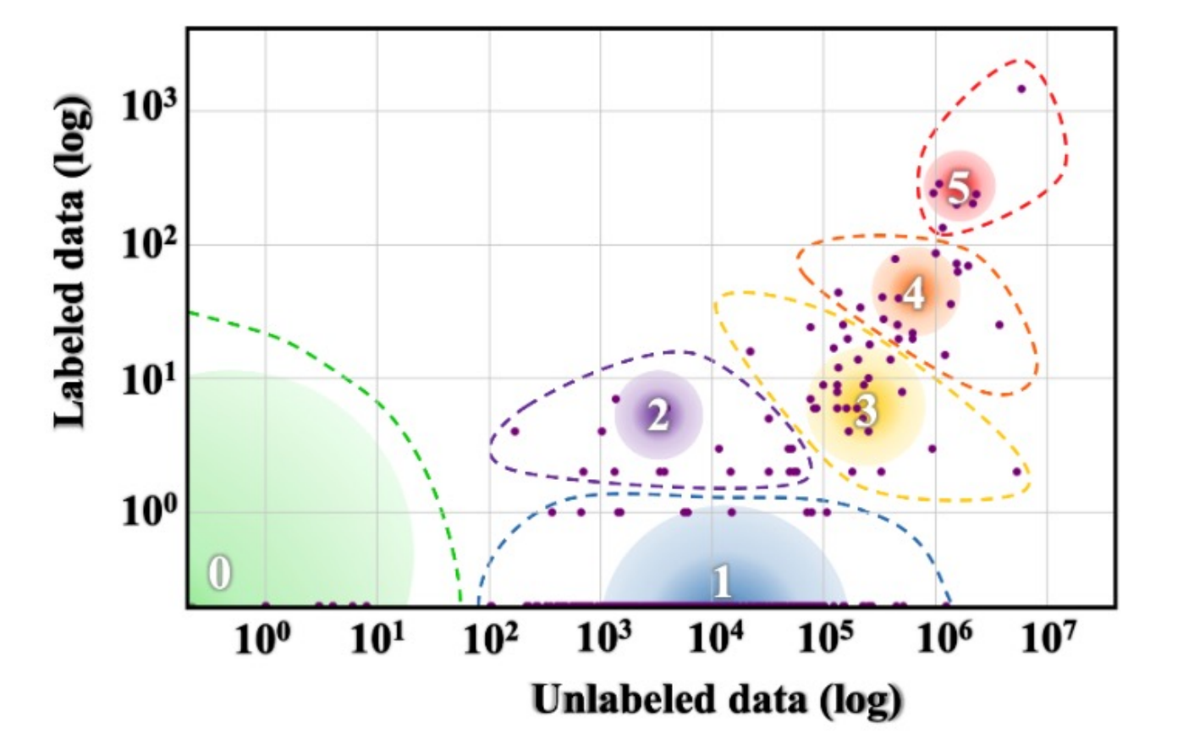


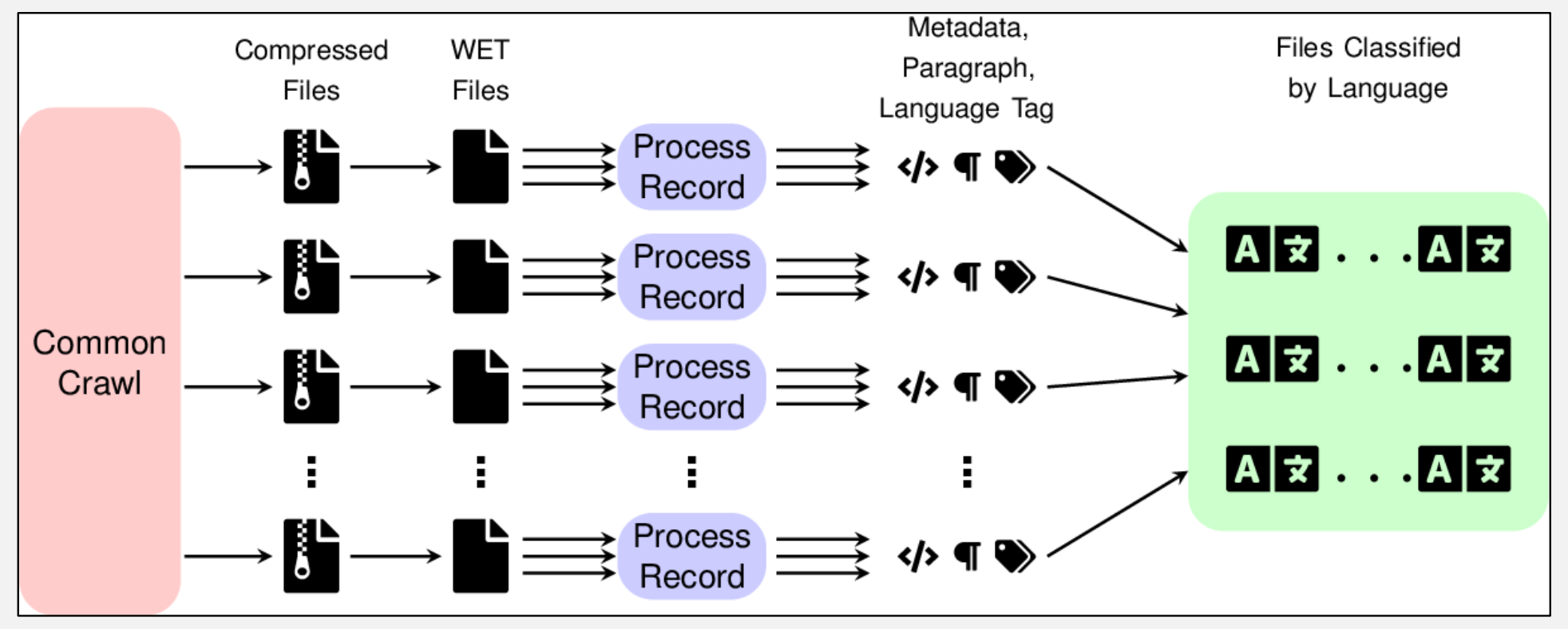
Figure 2: Language Resource Distribution: The size of the gradient circle represents the number of languages

However, there is no corpus available that

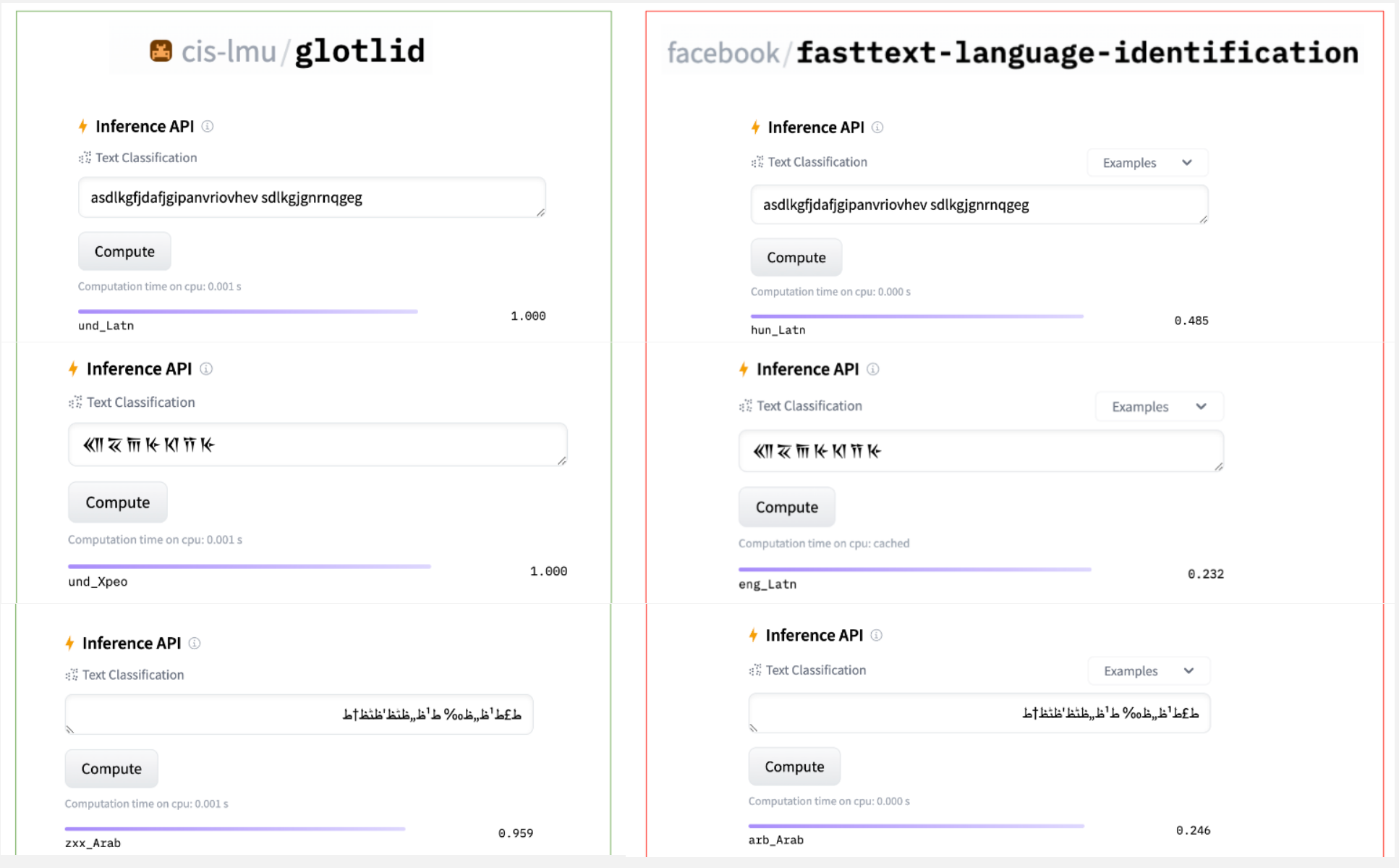
- covers a wide range of **minority** languages;
- is generated by an **open-source reproducible** Pipeline
- is rigorously **cleaned** from noise, making it trustworthy to use.

Pipeline

Base; The Ungoliant pipeline processes **CommonCrawl** WET files by extracting metadata, creating paragraphs, and tagging content with a **language identification (LID) model**. The output includes metadata, structured paragraphs, and language tags, with records grouped into files by language.



Language identification; LID is typically understood as a closed-set classification problem; most LID systems adopt this setup. However, since LID is inherently an open-set problem, processing web data always carries the risk of encountering "unknown" languages. We developed GlotLID v3, which covers nearly 2,000 labels, including major web noise labels and unseen writing systems.



cis-lmu / glotlid

Inference API
Text Classification
asdlkgfjdafjgipanvriovhev sdlkgjgnrnqgeg
Compute
Computation time on cpu: 0.001 s
und_Latn 1.000

Inference API
Text Classification
[illegible]
Compute
Computation time on cpu: 0.001 s
und_Xpeo 1.000

Inference API
Text Classification
[illegible]
Compute
Computation time on cpu: 0.001 s
zxx_Arab 0.959

facebook / fasttext-language-identification

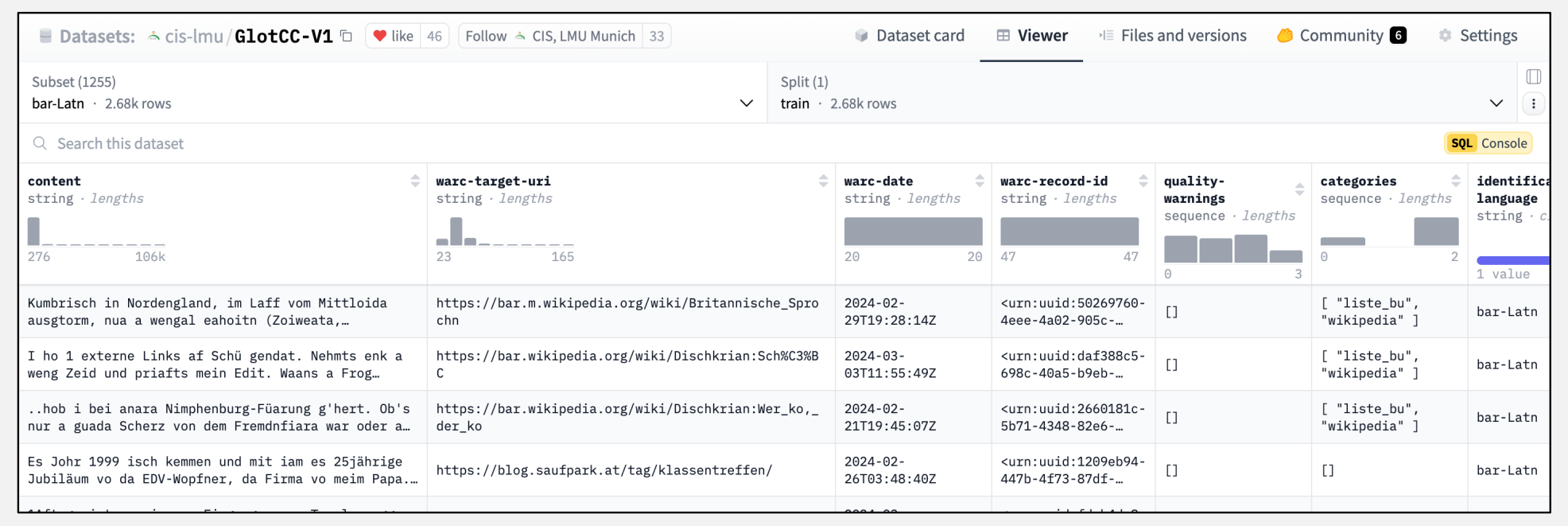
Inference API
Text Classification
Examples
asdlkgfjdafjgipanvriovhev sdlkgjgnrnqgeg
Compute
Computation time on cpu: 0.000 s
hun_Latn 0.485

Inference API
Text Classification
Examples
[illegible]
Compute
Computation time on cpu: cached
eng_Latn 0.232

Inference API
Text Classification
Examples
[illegible]
Compute
Computation time on cpu: 0.000 s
arb_Arab 0.246

Filters; We primarily use the cleaning process from OSCAR, MADLAD-400, GlotScript, and FineWeb. Since the LIDs are trained at the sentence level, we applied them to both sentences and entire documents, ensuring the majority of the data have consistent labels.

Self-audit quality review; Out of 653 audited languages, we find that, with a macro-average score of **0.93** and a median score of **1.0**, the data is in-language. There are still errors that neither the LID nor the filters captures. For example, repetitive n-grams in list-like content.



Datasets: cis-lmu / GlotCC-V1 ♥ like 46 Follow CIS, LMU Munich 33 Dataset card Viewer Files and versions Community 6 Settings

Subset (1255)
bar-Latn · 2.68k rows

Split (1)
train · 2.68k rows

Search this dataset SQL Console

content string · lengths 276 106k	warc-target-uri string · lengths 23 165	warc-date string · lengths 20 20	warc-record-id string · lengths 47 47	quality-warnings sequence · lengths 0 3	categories sequence · lengths 0 2	identifica language string 1 value
Kumbrisch in Nordengland, im Laff vom Mittloida ausgtorm, nua a wengal eahoitn (Zoiweata,…	https://bar.m.wikipedia.org/wiki/Britannische_Sprochn	2024-02-29T19:28:14Z	<urn:uuid:50269760-4eee-4a02-905c-…	[]	["liste_bu", "wikipedia"]	bar-Latn
I ho 1 externe Links af Schü gendat. Nehmts enk a weng Zeid und priafts mein Edit. Waans a Frog…	https://bar.wikipedia.org/wiki/Dischkrian:Sch%C3%BC	2024-03-03T11:55:49Z	<urn:uuid:daf388c5-698c-40a5-b9eb-…	[]	["liste_bu", "wikipedia"]	bar-Latn
..hob i bei anara Nimphenburg-Füarung g'hert. Ob's nur a guada Scherz von dem Fremdnfiara war oder a…	https://bar.wikipedia.org/wiki/Dischkrian:Wer_ko,_der_ko	2024-02-21T19:45:07Z	<urn:uuid:2660181c-5b71-4348-82e6-…	[]	["liste_bu", "wikipedia"]	bar-Latn
Es Johr 1999 isch kemmen und mit iam es 25jährige Jubiläum vo da EDV-Wopfner, da Firma vo meim Papa.…	https://blog.saufpark.at/tag/klassentreffen/	2024-02-26T03:48:40Z	<urn:uuid:1209eb94-447b-4f73-87df-…	[]	[]	bar-Latn

GlotCC vs other attempts (e.g., OSCAR)

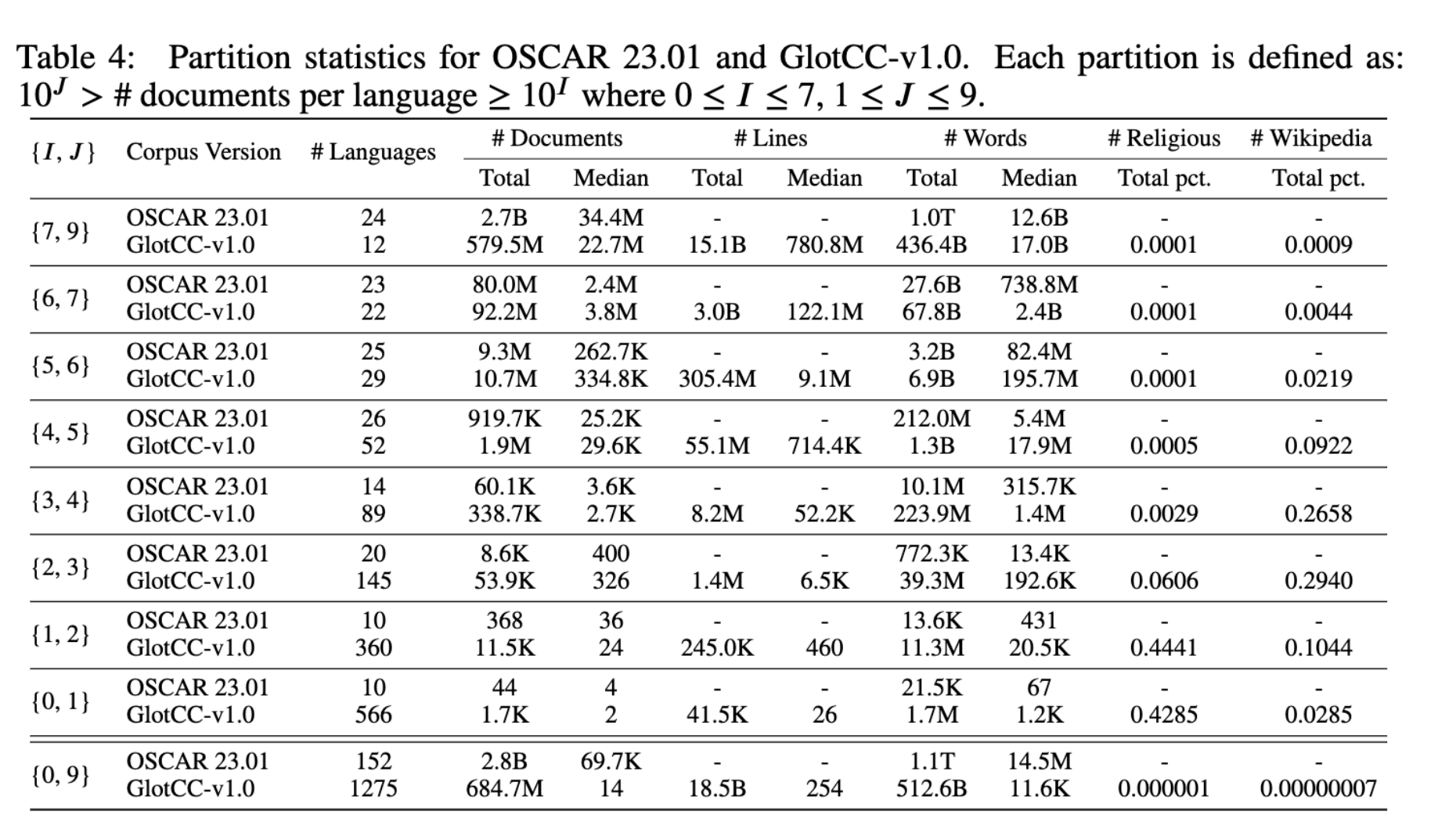
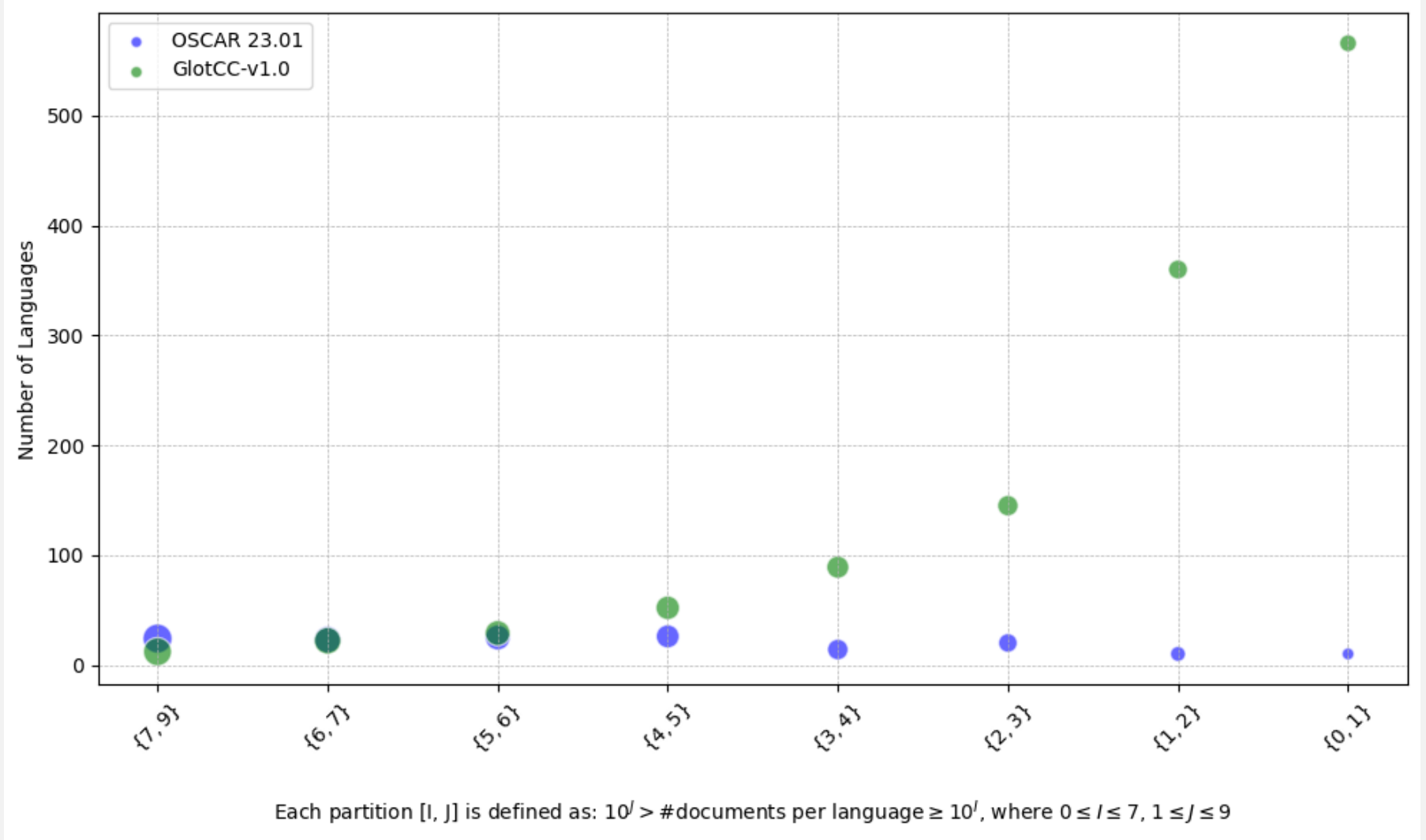


Table 4: Partition statistics for OSCAR 23.01 and GlotCC-v1.0. Each partition is defined as: $10^J >$ # documents per language $\geq 10^I$ where $0 \leq I \leq 7, 1 \leq J \leq 9$.

$\{I, J\}$	Corpus Version	# Languages	# Documents		# Lines		# Words		# Religious	# Wikipedia
			Total	Median	Total	Median	Total	Median	Total pct.	Total pct.
{7, 9}	OSCAR 23.01	24	2.7B	34.4M	-	-	1.0T	12.6B	-	-
	GlotCC-v1.0	12	579.5M	22.7M	15.1B	780.8M	436.4B	17.0B	0.0001	0.0009
{6, 7}	OSCAR 23.01	23	80.0M	2.4M	-	-	27.6B	738.8M	-	-
	GlotCC-v1.0	22	92.2M	3.8M	3.0B	122.1M	67.8B	2.4B	0.0001	0.0044
{5, 6}	OSCAR 23.01	25	9.3M	262.7K	-	-	3.2B	82.4M	-	-
	GlotCC-v1.0	29	10.7M	334.8K	305.4M	9.1M	6.9B	195.7M	0.0001	0.0219
{4, 5}	OSCAR 23.01	26	919.7K	25.2K	-	-	212.0M	5.4M	-	-
	GlotCC-v1.0	52	1.9M	29.6K	55.1M	714.4K	1.3B	17.9M	0.0005	0.0922
{3, 4}	OSCAR 23.01	14	60.1K	3.6K	-	-	10.1M	315.7K	-	-
	GlotCC-v1.0	89	338.7K	2.7K	8.2M	52.2K	223.9M	1.4M	0.0029	0.2658
{2, 3}	OSCAR 23.01	20	8.6K	400	-	-	772.3K	13.4K	-	-
	GlotCC-v1.0	145	53.9K	326	1.4M	6.5K	39.3M	192.6K	0.0606	0.2940
{1, 2}	OSCAR 23.01	10	368	36	-	-	13.6K	431	-	-
	GlotCC-v1.0	360	11.5K	24	245.0K	460	11.3M	20.5K	0.4441	0.1044
{0, 1}	OSCAR 23.01	10	44	4	-	-	21.5K	67	-	-
	GlotCC-v1.0	566	1.7K	2	41.5K	26	1.7M	1.2K	0.4285	0.0285
{0, 9}	OSCAR 23.01	152	2.8B	69.7K	-	-	1.1T	14.5M	-	-
	GlotCC-v1.0	1275	684.7M	14	18.5B	254	512.6B	11.6K	0.000001	0.00000007



From high-resource to low-resource, the size of each circle is proportional to the logarithm of the total number of documents.

References

- GlotLID: Language Identification for Low-Resource Languages, 2023
- Ungoliant: An Optimized Pipeline for the Generation of a Very Large-Scale Multilingual Web Corpus, 2021
- Towards a Cleaner Document-Oriented Multilingual Crawled Corpus, 2022
- MADLAD-400: A Multilingual And Document-Level Large Audited Dataset, 2023
- GlotScript: A Resource and Tool for Low Resource Writing System Identification, 2024